Tesis de maestría
Master in Management+Analytics
Universidad Torcuato Di Tella

Un nuevo enfoque de Privacidad de los Datos aplicado a resoluciones judiciales

En colaboración con el Juzgado Penal, Contravencional y de Faltas N°10 de la Ciudad Autónoma de Buenos Aires

Alumno: Lionel Alberto Barbagallo
Tutor: Pablo Roccatagliata

MAYO de 2021

UNIVERSIDAD TORCUATO DI TELLA

Resumen

Durante los últimos años ha ganado impulso el movimiento de gobierno abierto, que busca transparentar la administración, acercando el Estado a la ciudadanía, pero también abriendo repositorios de información con el fin de favorecer el avance del conocimiento en diversas áreas de investigación. En este contexto, diversos actores del poder judicial en la Argentina han implementado programas de apertura de datos, entre ellos el Juzgado Penal, Contravencional y de Faltas N° 10 de la Ciudad Autónoma de Buenos Aires. No obstante, estas iniciativas de democratización de los datos y del conocimiento chocan con un desafío: el de la privacidad de los individuos. Dada la gran expansión de Internet, el incremento en el acceso a nuevos y masivos conjuntos de datos y en el poder de cómputo y tecnología de manipulación de los datos; las técnicas y enfoques tradicionales de protección de la privacidad resultan insuficientes. Para superar estas amenazas a la privacidad en la implementación de su programa de datos abiertos, estaremos colaborando con el Juzgado N° 10 a fin de proponer una estrategia de apertura de datos capaz de garantizar elevados estándares de privacidad a los individuos. Al finalizar este trabajo, realizaremos una propuesta alternativa al flujo de trabajo actualmente utilizado por el Juzgado. Para ello, nos valdremos de una novedosa técnica llamada *Differential Privacy,* que permite llevar adelante la distribución de conjuntos de datos, pero controlando la cantidad de información privada que se deja filtrar.

Abstract

In recent years, the open government movement has gained momentum. This movement aims to make the administration transparent, bringing the State closer to the citizenry, but also opening information repositories in order to promote the advancement of knowledge in various research areas. In this context, various actors of the judiciary system in Argentina have implemented data opening programs, including Criminal, Misconduct and Misdemeanor Court No. 10 of the Autonomous City of Buenos Aires. However, these initiatives to democratize data and knowledge come up against a challenge: that of individual privacy. Given the great expansion of the Internet, the increase in access to new and massive data sets and in the computing power and data manipulation technology; traditional privacy protection techniques and approaches are insufficient. To overcome these threats to privacy in the implementation of its open data program, we will be collaborating with Court No. 10 in order to propose a data opening strategy capable of guaranteeing high privacy standards for individuals. At the end of this work, we will make an alternative proposal to the workflow currently used by the Court. To do this, we will use a new technique called Differential Privacy, which allows us to carry out the distribution of data sets, but controlling the amount of private information that is allowed to be leaked.

Prefacio

Este trabajo ha sido llevado adelante en colaboración del Juzgado Penal, Contravencional y de Faltas N° 10 de la Ciudad Autónoma de Buenos Aires. Los datos sobre los que trabajaremos, han sido provistos por el Juzgado y son de público acceso (aunque entre las recomendaciones del trabajo proponemos una nueva estrategia de difusión de los datos).

El cuerpo de esta tesis se complementa con el Git Hub del proyecto[1] , donde pueden consultarse los datasets originales e intermedios utilizados a lo largo del trabajo. Asimismo, allí puede accederse a los notebook's donde se presenta el procesamiento de los datos. Aunque en el cuerpo de la tesis hacemos una presentación de la metodología y criterios de tratamiento de los datos adoptada, alentamos al lector a consultar los notebook´s a fin de comprender los detalles de implementación. Para facilitar el seguimiento, los notebooks se hallan debidamente comentados.

[1] https://github.com/Lionelbarbagallo/Un-nuevo-enfoque-de-Privacidad-de-los-Datos-aplicado-a-expedientes-judiciales

Contenidos

Introducción. Gobierno Abierto y Privacidad.

Durante los últimos años ha ganado impulso el movimiento de gobierno abierto, que busca transparentar la administración, acercando el Estado a la ciudadanía, pero también abriendo repositorios de información con el fin de favorecer el avance del conocimiento en diversas áreas de investigación. En este sentido, las iniciativas de gobierno abierto se vinculan con el movimiento de datos abiertos, que busca hacer público el acervo de datos acumulado en archivos fragmentados, tanto del sector público, como privado, para favorecer el avance científico, social y tecnológico.

Política de datos abiertos en Juzgado Penal, Contravencional y de Faltas N°10 de la Ciudad de Buenos Aires

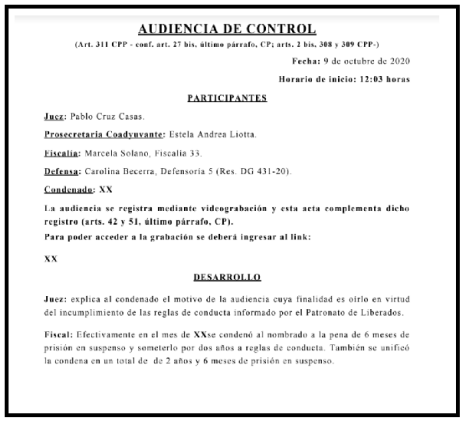
Desde 2012 Argentina adhiere a la Alianza para el Gobierno Abierto, una iniciativa global y multilateral que aboga por la implementación de políticas de transparencia de la gestión y rendición de cuentas hacia la ciudadanía (Jefatura de Gabinete de Ministros de Argentina, 2020). En este contexto, existen varias iniciativas en distintos niveles y sectores del Estado para implementar este tipo de políticas.

Dentro de este conjunto de iniciativas de gobierno abierto, el poder judicial no es la excepción. Ya sea encuadradas en normativas y leyes nacionales o provinciales, o por iniciativa individual de ciertos juzgados, existen en el país varias propuestas en esta dirección (que repasaremos con mayor detalle en la entrada 1.2). Entre ellas, destacan las iniciativas de los Juzgados N° 10 y N 13 de la Ciudad Autónoma de Buenos Aires[2]. En general, estas iniciativas apelan a dos estrategias de publicación de la información judicial:

1) Para favorecer la difusión de los actos de gobierno, se publican on - line todos los textos de sus resoluciones.
2) Asimismo, también se publican on - line datasets en formato tabular con información relativa a las causas, incluyendo datos respecto a las acusaciones, a los imputados, a las víctimas y las resoluciones. Con ello, los juzgados buscan acercar la información al público investigador y también publicitar sus actos de gobierno, promoviendo niveles de transparencia y cercanía con la ciudadanía.

[2] Estas no son las únicas iniciativas de apertura de datos en la justicia. Como se verá más adelante, existen encuadres nacionales, y algunos provinciales también.

Esquema I. Estrategias de publicación de la información actualmente utilizadas por el Juzgado N° 10



AUDIENCIA DE CONTROL

(Art. 311 CPP - conf. art. 27 bis, último párrafo, CP; arts. 2 bis, 308 y 309 CPP-)

Fecha: 9 de octubre de 2020

Horario de inicio: 12:03 horas

PARTICIPANTES

Juez: Pablo Cruz Casas.

Prosecretaria Coadyuvante: Estela Andrea Liotta.

Fiscalía: Marcela Solano, Fiscalía 33.

Defensa: Carolina Becerra, Defensoría 5 (Res. DG 431-20).

Condenado: XX

La audiencia se registra mediante videograbación y esta acta complementa dicho registro (arts. 42 y 51, último párrafo, CP).

Para poder acceder a la grabación se deberá ingresar al link:

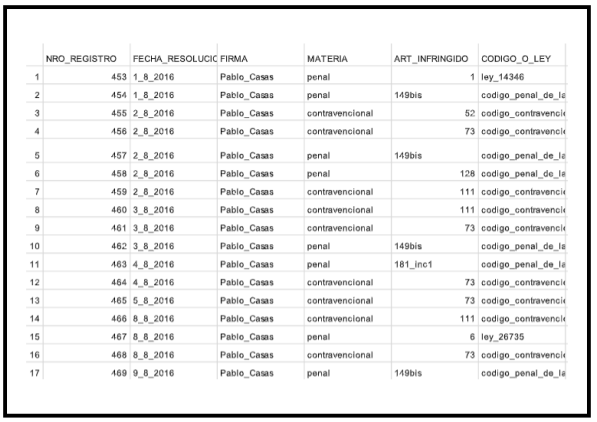
XX

DESARROLLO

Juez: explica al condenado el motivo de la audiencia cuya finalidad es oírlo en virtud del incumplimiento de las reglas de conducta informado por el Patronato de Liberados.

Fiscal: Efectivamente en el mes de **XX**se condenó al nombrado a la pena de 6 meses de prisión en suspenso y someterlo por dos años a reglas de conducta. También se unificó la condena en un total de de 2 años y 6 meses de prisión en suspenso.

Texto de las Resoluciones



	NRO_REGISTRO	FECHA_RESOLUCIC	FIRMA	MATERIA	ART_INFRINGIDO	CODIGO_O_LEY
1	453	1_8_2016	Pablo_Casas	penal	1	ley_14346
2	454	1_8_2016	Pablo_Casas	penal	149bis	codigo_penal_de_la
3	455	2_8_2016	Pablo_Casas	contravencional	52	codigo_contravencio
4	456	2_8_2016	Pablo_Casas	contravencional	73	codigo_contravencio
5	457	2_8_2016	Pablo_Casas	penal	149bis	codigo_penal_de_la
6	458	2_8_2016	Pablo_Casas	penal	128	codigo_penal_de_la
7	459	2_8_2016	Pablo_Casas	contravencional	111	codigo_contravencio
8	460	3_8_2016	Pablo_Casas	contravencional	111	codigo_contravencio
9	461	3_8_2016	Pablo_Casas	contravencional	73	codigo_contravencio
10	462	3_8_2016	Pablo_Casas	penal	149bis	codigo_penal_de_la
11	463	4_8_2016	Pablo_Casas	penal	181_inc1	codigo_penal_de_la
12	464	4_8_2016	Pablo_Casas	contravencional	73	codigo_contravencio
13	465	5_8_2016	Pablo_Casas	contravencional	73	codigo_contravencio
14	466	8_8_2016	Pablo_Casas	contravencional	111	codigo_contravencio
15	467	8_8_2016	Pablo_Casas	penal	6	ley_26735
16	468	8_8_2016	Pablo_Casas	contravencional	73	codigo_contravencio
17	469	9_8_2016	Pablo_Casas	penal	149bis	codigo_penal_de_la

Set de Datos Tabulados

Pero estas iniciativas hacia la apertura y circulación de la información, tienen por delante un desafío importante, el de la **privacidad**. A medida que la vida social se fue volcando hacia el ámbito digital, cada vez mayores volúmenes de información personal fueron recolectados y distribuidos por diversos sistemas. Compartir la información existente en tales repositorios, puede, y en muchos casos lo hace, constituir una amenaza a la privacidad.

En el contexto actual, de fácil acceso a la información, gran poder de cómputo, y desarrollo de nuevas tecnologías, los riesgos a la privacidad de los individuos son crecientes. Las técnicas tradicionales de anonimización resultan insuficientes para proteger la identidad de los individuos incluidos en un conjunto de datos. A partir de unos pocos datos, un oponente puede orquestar un ataque de identificación, con gran probabilidad de éxito. Por otro lado, la amenaza de la reconstrucción es un riesgo cada vez mayor. Se ha comprobado, que a partir de cierta información, en teoría despojada de datos privados, se puede reconstruir otra mucho más comprometedora. Por ejemplo, a partir de estadísticas agregadas, es posible reconstruir los registros individuales. A partir de ciertos registros médicos y biológicos, se puede lograr una reconstrucción de la imagen facial muy acertada. Lo que estos ejemplos ponen en evidencia, es que la masiva virtualización de la vida social y el vertiginoso desarrollo tecnológico suponen nuevos riesgos y desafíos a la privacidad[3].

[3] En el apartado 1.1.2 se hace un repaso de estos riesgos. Para una visión de contexto, recomendamos ver el capítulo 1 de Dwork & Roth (2014).

Frente a esta realidad, las estrategias de apertura de datos implementadas por los Juzgados 10 y 13 no escapan a los riesgos a la privacidad de los individuos y filtrado de información sensible. A pesar de que los datasets tabulares publicados por los Juzgados no contienen datos directamente identificatorios (de ahora en más PII, por sus siglas en inglés), como nombres, direcciones o teléfonos, estos conjunto de datos son susceptibles a un ataque de identificación capaz de exponer las identidades de los víctimas y agresores y exponer gran cantidad de información privada[4].

Objetivos del trabajo: colaboración con el Juzgado N° 10 en su estrategia de datos abiertos

A lo largo de este trabajo, estaremos colaborando con el Juzgado N° 10 para presentar una propuesta superadora, capaz de alcanzar elevados criterios de privacidad para su proyecto de apertura de datos. **Como salida de este proyecto, presentaremos un flujo de trabajo y recomendaciones de cara a la implementación de un futuro *Minimum Viable Product (MVP*) por parte del Juzgado, capaz de garantizar rígidos estándares de seguridad en la publicación de la información que actualmente se distribuye en formato estructurado (no abordaremos el problema de la publicación de los textos de las resoluciones)**. Para ello, nos basaremos en el marco conceptual de ***Differential Privacy***, un marco conceptual que permite la publicación de información minimizando el goteo de datos personales.

Estructura del trabajo

Este trabajo se estructura en una introducción, dos partes y una conclusión. En la parte I, presentaremos todos los aspectos teóricos y técnicos de la disciplina. Asimismo, examinaremos en detalle las necesidades del Juzgado en lo que son sus estrategias de publicación de datasets, y las re - evaluaremos en función de los aportes teóricos introducidos. En la parte II del trabajo, presentaremos la implementación de la solución propuesta en base a los conjuntos de datos disponibles. Finalmente, en las conclusiones recapitulamos sobre los aportes del trabajo y presentamos reflexiones de cara al futuro.

[4] El *National Institute of Standards and Technology* [NIST] (2007) de los Estados Unidos define los PII "como cualquier información sobre un individuo en poder de una agencia, incluyendo (1) cualquier información que pueda ser usada para distinguir o reconstruir la identidad de un individuo, como su nombre, número de la seguridad social, fecha y lugar de nacimiento, apellido materno, registros biométricos; y (2) cualquier otra información que pueda vincularse a un individuo, tal como información médica, educacional, financiera o de empleo (*Trad. del A.*)."

La parte I del trabajo se compone de cinco secciones. En la primera sección haremos una presentación de los debates en torno a la privacidad y sus desafíos en el contexto de la era digital. En la segunda sección haremos un repaso de los puntos más salientes del movimiento de justicia abierta y de la normativa existente. Al finalizar esta sección revisaremos la problemática de apertura de los datos del Juzgado N° 10 a la luz de los debates presentados hasta el momento. La tercera sección incluye una presentación detallada de lo que es *Differential Privacy,* la metodología de tratamiento de los datos que utilizaremos. La cuarta sección hace un repaso por algunas implementaciones de *Differential Privacy* en la práctica y los diversos dominios en los que se puede utilizar. La quinta sección hace un repaso por el ecosistema tecnológico disponible para aplicar *Differential Privacy*.

La parte II se compone de seis secciones. En la primera sección se analizarán las estrategias de alto nivel disponibles para la aplicación de mecanismos de *Differential Privacy* en función de los datasets a trabajar. En la segunda sección se realizará un análisis exploratorio de los datos y se establecerán algunos criterios básicos de tratamiento de los mismos. En la tercera sección se presentará la ingeniería de atributos realizada sobre el dataset. En la cuarta sección se definirá la implementación de los mecanismos de *Differential Privacy* . En la quinta sección se analizarán diversos factores que afectan a la *performance* de los mecanismos. Finalmente, en la sexta sección se hará un repaso con las sugerencias y el flujo de trabajo propuesto.

Parte I.

Marco Teórico

Como se comentó en la introducción del trabajo, la apertura de datos en cualquier esfera - incluida la apertura de registros gubernamentales o judiciales - involucra riesgos a la privacidad. Riesgos que en algunos casos son conocidos hace tiempo y riesgos que han surgido - o se han intensificado - en el contexto de la revolución en la informática y las comunicaciones de las últimas décadas. Abordar estos problemas requiere el refinamiento de un marco teórico para conceptualizar con precisión los desafíos a los que nos enfrentamos ¿Qué es la privacidad? ¿Cómo se ve afectada por la acelerada virtualización de la vida social? ¿Qué riesgos existen - si los hay - en las estrategias de apertura de datos implementadas por el Juzgado N°10? Y sobre todo ¿Cómo podemos actuar frente a tales amenazas? ¿Qué herramientas tenemos a disposición para resguardar la privacidad en este entorno desafiante y cambiante?

A lo largo de la parte I de la tesis estaremos introduciendo el marco teórico fundamental para abordar estas preguntas y poder trabajar sobre estos problemas, para en la parte II, poder presentar una solución técnica capaz de alcanzar los objetivos propuestos.

1.1. Problemas en torno a la privacidad

1.1.1. ¿Qué entendemos por privacidad?

Sin dudas la definición de privacidad es un punto complejo, que descansa en consideraciones filosóficas, morales y sociales. Es difícil dar una única y certera definición de privacidad. En cierto grado, cada individuo tiene una idea más o menos diferente respecto a los eventos e información que considera privados ¿Es un nombre y apellido información privada? ¿Es una cara en una imagen información privada? ¿Es un *curriculum vitae* información privada? Aunque estas parecen preguntas que quizás a nivel individual pueden ser fáciles de responder, veremos que en realidad no existe una única respuesta a ellas.

Helen Nissenbaum (2010) propone una serie de definiciones para pensar el problema. En general, estamos acostumbrados a pensar la privacidad en términos binarios. O bien algo es privado, o no lo es. Si bien reconocemos que a veces es difícil encasillar todas las situaciones en cada una de estas definiciones (como es el caso de las preguntas de arriba), lo más común es que tratemos de hacerlo. Esto es mucho más sencillo para información y situaciones sobre las que existe consenso: son privados datos relativos a menores, orientación sexual, religiosa, orientación política y registros médicos. No casualmente, donde existe este consenso, la ley ha reglado sobre estos aspectos. Es bastante común que la ley incluya referencias explícitas a este tipo de datos. Pero con excepciòn de estos casos, en general no existe una ùnica idea respecto a lo que es la privacidad.

El primer paso para avanzar en la construcción de una definición màs compleja de privacidad, pasa por desembarazarse del esquema binario. Según Nissenbaum (2010), **la información no es ni deja de ser privada por razones intrínsecas a la misma, sino que la consideración de privacidad depende del contexto situacional y de cómo fluye esta información**. Así, el mismo *bit* de información puede ser en un contexto privado y en otro no. Por ejemplo, si un individuo está en su jardín tomando sol, es consciente de que está a la vista de sus vecinos. No por ello ha renunciado a la privacidad. Este individuo sabe que está dejando fluir cierta información, no obstante, sabe que en este caso, está limitado a los pocos transeúntes que pueden pasar delante de su casa. En este caso hipotético, el individuo acepta este flujo de información. No obstante ¿Cambiaría la situación si alguien le tomase fotografías y las publicase on-line? Una lectura simplista argumentaría que una vez que la persona decidió tomar sol en su jardín, y sabiendo que está a la vista de todos sus vecinos, renunció a la expectativa de privacidad. Es decir, si el individuo decidió que no tiene

problemas con que lo vean tomando sol sus vecinos, el argumento se extiende a que dicha persona tampoco debería tener problemas con que lo viesen tomando sol millones de personas de forma on-line. Pero, en realidad, esta persona accedió a que (en el peor de los casos), sus vecinos lo vean tomando sol. Sólo sus vecinos, y nadie más. Es decir, aceptó que cierta información (su imagen tomando sol), fluya de una manera específica y anticipada (sólo hacia sus vecinos). Nótese que esto escapa a la definición binaria. Propuesta de este modo, la definición de privacidad pasa por garantizar y respetar un flujo deseado de la información, no por una característica inherente a la misma.

Veamos otro ejemplo mundano. En una reunión de amigos, alguien puede hacer un chiste. En dicho contexto, no tiene consecuencias mayores. Ahora, imaginemos que alguien filma de forma oculta esta situación y el video se hace viral. Al margen del contenido del chiste (supongamos que no estuvo fuera de cánones aceptables y políticamente correctos), esta situación no fue anticipada por el individuo y constituye una violación a su expectativa de privacidad. Pero, lo que es privado no es el chiste en sí. En el contexto de la reunión con amigos, la circulación de esta información fue positiva, fue deseada. Ahora, imaginemos que este video llega a los jefes de esta persona en la oficina. Puede que esto le traiga (o no), algún problema o incomodidad. Este es un flujo no deseado de la información. Lo que vemos en este otro ejemplo, es que más que la información en sí, lo que constituye una violación a la expectativa de privacidad es el sentido en que circula. Si la información circula en formas no anticipadas ni aceptadas por los individuos, constituye una violación a su expectativa de privacidad.

El principio de privacidad contextual nos lleva a tener en consideración principalmente los flujos de información - no tanto la información en sí misma. A diferencia de lo que un planteo binario de la privacidad postularía, bajo este enfoque no existe información puramente privada (ni pública), sino que ello debe definirse en función de la expectativa de circulación de la misma que tienen los individuos.

1.1.2. Riesgos a la privacidad en la era del '*big data*'

Los riesgos que implica la creciente digitalización y virtualización de la vida social no representan necesariamente un nuevo tipo de amenaza a la privacidad. Sin dudas, muchos de los problemas que se evidencian en la era digital, existían también en la era analógica. En realidad, la era digital abre la puerta a un aumento cuantitativo de las amenazas a la privacidad, pero, *per se,* no crea nuevas formas de amenazas. Si entendemos una vulneración a la privacidad como la circulación no deseada (y probablemente perjudicial) de

información, veremos que este riesgo siempre existió. Aún antes de la era digital, la información podía, a través de medios analógicos, circular de formas no deseadas, suponiendo graves riesgos a la privacidad y perjuicios a los individuos. Basta con considerar el impacto de la imprenta o de la fotografía para darse cuenta cómo un medio analógico puede suponer importantes vulnerabilidades a la privacidad (Nissenbaum, 2011). Por otro lado, el aparato de vigilancia estatal, muchas veces tan señalado como uno de los grandes temores de la revolución tecnológica, también funcionó en etapas previas, valiéndose de otros recursos técnicos (Bartolucci, 2017; Cornwall, 1992). Estos ejemplos ponen de relieve que la amenaza a la privacidad y a los individuos, no se limita únicamente al medio por el que circule la información, y en cambio, se vincula principalmente con los principios sociales, políticos, económicos y morales que guían su recopilación, circulación y utilizamiento posterior.

La novedad que presenta la era digital, y con ella el advenimiento de la vida virtual, la eclosión de Internet y de la capacidad de cómputo, en principio no es cualitativa, sino cuantitativa (Nissenbaum, 2011). Las últimas décadas han presenciado un crecimiento exponencial en la generación y digitalización de datos, en la circulación y en la capacidad de cómputo. Aunque la naturaleza de las amenazas a la privacidad es la misma que la que existía bajo un paradigma de vida analógico, esto que a primera vista aparece como una diferencia cuantitativa, en realidad se transforma en una diferencia cualitativa. El salto cuantitativo es tan grande, que por sí mismo, modifica la naturaleza de las relaciones sociales y las prácticas dominantes en ciertos entornos. Eso se conoce como ‘disrupción digital’ (Baiyere & Hukal 2020). Y aunque la lógica de los riesgos a la privacidad siguen siendo los mismos a los que existían en la era analógica, en esta nueva etapa, donde la vida se vuelve virtual e hiperconectada, la posibilidad de flujos inesperados y potencialmente dañinos de información, crece exponencialmente. Estas transformaciones modifican radicalmente el modo en que debe pensarse la privacidad en el contexto digital. Situaciones que en un contexto analógico no suponían amenazas, en este nuevo entorno deben reevaluarse.

Estas nuevas amenazas a la privacidad emanan de varios frentes. Primero, se generan y recopilan datos a gran escala. Esto es lo que nos referimos previamente como ‘virtualización de la vida social’. Esto involucra a los exponentes más evidentes, como las redes sociales, pero también al uso de todo tipo de plataformas y servicios on-line, la Internet de las cosas, dispositivos móviles y sistemas de vigilancia en la vía pública y privada. Alcanza esta breve enumeración para darse cuenta de todos los puntos de contacto con el ámbito virtual. Ya no se trata de un contacto ocasional, sino que ellos son la norma. En segundo lugar, Internet

abre la puerta a una velocidad y facilidad en la circulación de la información sin precedentes. Para que todo lo anterior pueda funcionar, se vale de la interconectividad de sistemas a través de Internet. Sistemas, dispositivos e individuos tienen acceso de forma casi instantánea a una gran cantidad de datos. De este modo, la posibilidad real de controlar los flujos de información se hace más escurridiza[5]. Lo que viene a complejizar este escenario es la capacidad de cómputo disponible, y la existencia de algoritmos capaces de explotar eficientemente estos volúmenes de datos. No sólo hay una gran cantidad de información circulando, de forma no prevista, sino que resulta económico organizar la explotación a gran escala de la misma (Nissenbaum, 2011).

1.1.3. Ataques sobre la privacidad

Si bien la era digital supone algunos riesgos que resultan evidentes y hasta obvios, otros no lo son. En general, son los riesgos más evidentes los que alarman al público. Por ejemplo, que una aplicación de correos electrónicos o mensajería sea vulnerada, que mensajes o imágenes comprometedoras lleguen a destinatarios no intencionados, o que un sistema de videovigilancia rastree nuestros movimientos. El mismo reparo existe frente a la información que se vuelca en aplicaciones, plataformas y dispositivos móviles o sensores de Internet de las cosas. Lo que tienen en común estas amenazas, es que involucran la circulación no deseada de la información. Si bien todos estos son riesgos reales, existen otros - no necesariamente más graves - pero sí menos conocidos, que constituyen amenazas latentes. **Estos riesgos se originan mayoritariamente en la posibilidad de extraer, a partir de datos que parecían 'inocentes' o carentes de otro tipo de valor, nueva información[6]**. Un ejemplo bastará para comprender esta posibilidad. Imaginemos que vamos a realizarnos una tomografía cerebral. Luego, dicha imagen - en teoría despojada de toda información personal - es publicada en un dataset abierto sobre enfermedades neurológicas con el fin de favorecer el avance científico. Nosotros, sabiendo que sólo se publicará la tomografía sin ninguna otra referencia sobre nuestra identidad, aceptamos de buena gana su publicación, ya que en principio no involucra riesgos a la privacidad y favorecerá el avance científico. Tal como está planteada la situación, estamos aceptando este flujo de información. No obstante, y aquí radica la problemática emergente, lo que no sabíamos en su momento, pero hoy sí se sabe, es que a partir de una tomografía computada se puede reconstruir con asombrosa

[5] Por ejemplo, a partir de los perfiles públicos de Facebook es posible inferir muchos datos privados del individuo, como orientación sexual, género, edad, entre otros (Kosinski, Stillwell & Graepel, 2013).

[6] Esto incluye por ejemplo al denominado problema de *bundling* presentado por Bluemke *et. al.* (2020). Este problema se quiere transmitir cierta información, pero esta viene intrínsecamente vinculada con otra que queremos mantener privada. Un caso (probablemente inofensivo) de esto se daría en el contexto de una videollamada, donde uno quiere que los demás participantes vean nuestra cara, pero no el fondo detrás nuestro (por eso en general las aplicaciones ofrecen la posibilidad de reemplazarlo).

precisión, el rostro de la persona. Y ya con la imagen facial, no sería muy difícil encontrar a qué persona corresponde, por ejemplo utilizando la búsqueda por imágenes de Google (Peng *et. al.*, 2021; Kolata, 2019). Aunque inicialmente creíamos que estábamos compartiendo una imagen sin mayores implicancias, esta terminó revelando muchísima información más de la que esperábamos, convirtiéndose en un una gran vulnerabilidad de nuestra privacidad. Este es el **riesgo de reconstrucción**[7]. A partir de cierta información, se logra reconstruir otra. A través del análisis de una imágen, se logran conocer otro tipo de datos.

Ejemplos como el anterior son variados. Kosinski & Wang (2018) lograron identificar la orientación sexual de un conjunto de individuos a partir de sus rasgos faciales. Algo similar a lo que ocurre con una tomografía ha sido probado para ciertos segmentos de la secuencia de ADN, que permiten la reconstrucción de la imagen facial de la persona (Lippert *et al.*, 2017). Saliendo del ámbito del análisis de imágenes, donde las redes neuronales han hecho avanzar mucho el campo, existen otros tipos de ataques de reconstrucción. Por ejemplo, un ejercicio interno del *US. Census Bureau* pudo reconstruir, a partir de los datos publicados del censo normeamericano de 2010, una proporción importante de los registros originales (*National Academies of Sciences, Engineering, and Medicine*, 2021). En otro dominio, Bolot & Zang (2011) han podido des anonimizar datasets de llamadas al servicio de emergencias de los Estado Unidos en base a los datos de geolocalización incluidos en los registros. Sus conclusiones son enfáticas: “Nuestro estudio demuestra que la publicación o divulgación de localizaciones anonimizadas probablemente pueda significar riesgos a la privacidad (...)” [Trad. del A.] (2011:11). El común denominador de estas situaciones es que información que se creía incapaz de revelar otros datos - deliberadamente o no - termina haciéndolo.

Este riesgo de reconstrucción da lugar de forma casi automática al **riesgo de la identificación**[8]. Habiendo recuperado información de cierto conjunto de datos, un atacante podría descubrir la identidad de los individuos referidos. Este es el caso que tratamos en el ejemplo de la tomografía. A partir de la misma, se puede reconstruir la imagen facial y con ella, identificar a la persona. Lo mismo vale para otros tipos de datos. Es posible que conjuntos de datos que por separado no tuviesen mayor relevancia, ni se esperara que significasen una vulnerabilidad a la privacidad, terminen siéndolo. Este es el famoso caso

[7] No todos los ataques de reconstrucción refieren a datos no estructurados o imágenes. Abowd, Garfinkel, & Martindale, (2019) han presentado un análisis detallado de riesgos de este tipo sobre conjuntos de datos tabulares.

[8] Dwork & Roth (2014) explican cómo a partir de un conjunto de datos anónimo, el mismo puede ser vinculado con un conjunto de datos no anónimo y proceder a la identificación de los individuos involucrados. Ellos se refieren a este ataque como *linkage attack*.

del concurso de Netflix de 2006. En dicho año, la empresa de streaming lanzó un concurso para mejorar el sistema de recomendación que utilizaba la plataforma. Para resolver este desafío, compartió un dataset conteniendo las puntuaciones de películas realizadas por los usuarios. En teoría, no debería haber problemas con ello, pués no se publicaron datos personales, sino únicamente los id´s de los usuarios y los ratings de las películas. No obstante, en base a este conjunto de datos, un equipo de la Universidad de Texas logró re identificar a los usuarios. Para ello, cruzó los registros publicados por Netflix con calificaciones de películas del sitio web Imdb. En base a los reviews que los usuarios dejaron en en este sitio, los investigadores lograron hallar patrones similares dentro del conjunto de usuarios del dataset de Netflix, con los que pudieron vincular ambos conjuntos de datos y exponer la identidad de los usuarios de Netflix (y sus preferencias cinematográficas) (Narayanan & Shmatikov, 2006).

En este escenario, las técnicas tradicionales de protección de la privacidad resultan insuficientes. Por ejemplo, la publicación de datasets estructurados con información relativa a individuos, puede dar lugar a ataques de identificación. Incluso, si estos conjuntos de datos no contuviesen en sí mismos datos identificatorios, un ataque de este tipo es posible, como revela el caso de Netflix. Tradicionalmente, se pensaba que un dataset despojado de atributos identificatorios (PII), era suficiente para proteger la privacidad de los individuos. No obstante, este tipo de información es altamente reveladora, aunque se exponga codificada o hasheada. Si el conjunto de datos mantiene la 'señal estadística' de los registros, no importa que su información se halle codificada, un oponente siempre podrá realizar un ataque de identificación (Li *et. al*., 2017).

Frente a este contexto de nuevos riesgos y potenciales ataques, la comunidad científica y académica ha comenzado a proponer soluciones para intentar resguardar la privacidad. Ellas se conocen como *Private Enhancing Technologies* (PET) y se hallan en numerosos estados de avance. Las más evolucionadas hasta el momento son las técnicas de *Secure Computation*, *Federated Learning* y *Differential Privacy* (Bluemke *et. al.* 2021). En este trabajo, estaremos haciendo uso de *Differential Privacy* (de ahora en más DP), que es una técnica para proteger diversos tipos de conjuntos de datos de ataques identificatorios. El fundamento de los métodos de *Differential Privacy* es la inclusión de ruido en los registros, a fin de distorsionar la 'señal estadística' a la que antes nos referíamos y evitar la posible identificación de los individuos.

1.1.4. Las leyes de protección de la privacidad

Los estados nacionales no se han quedado al margen de la irrupción de la era digital y han establecido diversas regulaciones, tratando de preservar sobre todo el derecho a la privacidad y la propiedad de los datos. La legislación más moderna y amplia a nivel mundial es la *General Data Protection Regulation* (GDPR) que está en vigor desde 2018 en la Unión Europea. Esta ley tiene como objetivo proteger los datos personales, y establece que todo uso de datos personales por parte de terceros debe contar con consentimiento informado. Asimismo, establece la propiedad sobre los datos personales. Un individuo puede exigir a una empresa u organismo que le informen sobre la existencia de datos personales en sus registros, y puede exigir que se entreguen o eliminen. En la Argentina existe una ley similar, la 25.326 que data del año 2000 y a grandes rasgos, presenta las mismas definiciones y salvaguardas. En los Estados Unidos, no existe una ley tan amplia de protección de datos personales, pero ciertos sectores tienen regulaciones específicas. El sector de la salud está especialmente regulado y se encuentra bajo la órbita de la *Health Insurance Portability and Accountability Act* (HIPAA). Esta ley establece cómo debe manejarse la recopilación, circulación y almacenamiento de datos médicos y protege la privacidad de los datos personales dentro de los registros médicos.

Un punto a remarcar de estas leyes, es que parten de definiciones ambiguas y limitadas de lo que representa un dato personal. En general, restringen la definición de datos personales a atributos identificatorios (PII). O si se refieren de una forma más amplia a los datos personales, lo hacen de forma ambigua. Según la Unión Europea, "los datos personales son cualquier información relacionada con un individuo, ya sea que se refiera a su vida privada, profesional o pública. Puede ser cualquier cosa desde un nombre, domicilio, foto, dirección de correo electrónico, detalles bancarios, publicaciones en sitios web de redes sociales, información médica o la dirección IP de una computadora" (Comisión Europea, 2012). En tanto, la ley argentina define como datos personales: “información de cualquier tipo referida a personas físicas o de existencia ideal determinadas o determinables” (LPDP, 2000). En la práctica, este tipo de definiciones resultan complejas de aplicar. Bajo la definición estrecha, una gran cantidad de información no es definida como datos personales, pero bajo la definición amplia, todo puede terminar siendo un dato personal.

A pesar de estas limitaciones, estas leyes garantizan un umbral mínimo de seguridad y expresan la voluntad de legislar sobre el problema. No obstante, las soluciones que emanan de los requerimientos de las leyes, son todavía vulnerables a diversos tipos de ataques, muchos de los que hemos repasado anteriormente. Si bien la ley protege la identidad de los

individuos prohibiendo la publicación de PII (salvo consentimiento del individuo), conjuntos de datos anonimizados pueden aún revelar datos privados. Los mismo pasa con tipos de información más complejos, como muestran los casos presentados en las secciones precedentes.

Aunque estas leyes marcan una dirección en la preservación de la privacidad, muchas de las definiciones en las que se basan están inspiradas en reglas de juego de un mundo analógico, más que digital. En este sentido, tanto la definición de lo que se considera un dato personal, como las técnicas de manipulación aceptadas, resultan insuficientes a la luz de los riesgos a la privacidad existentes. Por ello, resulta conveniente re pensar estas definiciones en un marco conceptual más amplio, tal como el propuesto por Nissenbaum, de privacidad en contexto, y considerando nuevos tipos de riesgos no previstos en la legislación.

Una estrategia de apertura de datos, especialmente si involucra datos basados en personas, debe no sólo atenerse a las restricciones que emanan de las leyes existentes, sino que debe tener en cuenta la complejidad de la noción de lo 'privado' y lo 'personal', dados los problemas que tienen definiciones estáticas como las expuestas en las leyes y regulaciones existentes. Asimismo, una estrategia de datos abiertos debe considerar también los riesgos emanados de la progresiva virtualización de la vida social y revolución técnica. En este sentido, las leyes de protección de los datos personales deben tomarse como una base mínima (e insuficiente) para comenzar a definir un estándar adecuado de privacidad para estos proyectos.

1.2. Privacidad y acceso a los datos en la justicia

1.2.1. El acceso a la información judicial en el mundo y en la Argentina

El movimiento de justicia abierta se incluye en la dinámica más amplia de gobierno abierto. En los últimos años, ha ido ganando impulso la demanda de transparencia en los distintos ámbitos de la vida pública (Basterra, 2018). Demandas impulsadas desde la sociedad civil, ONG´s e incluso desde ciertos sectores políticos han abogado por mayor publicidad de los actos de gobierno. La transparencia en los actos de gobierno lleva a mayor control de la ciudadanía de la cosa pública, promoviendo mayor eficiencia y ejerciendo un control sobre potenciales delitos en el ejercicio de la función pública. Pero la transparencia no solo actúa como garantía contra la corrupción, sino también como puntal de un desarrollo colaborativo, donde el escrutinio y la participación de parte de la ciudadanía puede llevar a mejorar procesos y el surgimiento de nuevas propuestas (Jefatura de Gabinete de Ministros de Argentina, 2020; Gómez Zavaglia, 2020).

El derecho al acceso a la información pública ha sido reivindicado a nivel teórico y legal en numerosas instancias. La Organización de Estados Americanos promueve la adopción de leyes de gobierno abierto (OEA, 2009). Entre sus consideraciones, señala que el acceso a la información pública es un derecho humano fundamental y que incluye a toda la información en poder del sector público. En la Argentina, este principio se halla aceptado y existe una ley de acceso a la información pública. La Ley N° 27275 garantiza el derecho de acceso a la información pública dentro de la esfera nacional. A nivel provincial, existen leyes propias en el mismo sentido[9]. En la Ciudad Autónoma de Buenos Aires, el acceso a los datos públicos se halla consagrado por la ley 104 de 1998.

Este movimiento de gobierno transparente y acceso a la información pública incluye también a la justicia (Gómez Zavaglia, 2020). Como uno de los poderes de la república, también existe expectativa de transparencia en sus actos de gobierno. En este sentido, existen movimientos tanto a nivel internacional, como local que impulsan este tipo de prácticas. A nivel mundial, existe una alianza de institutos especializados en problemas judiciales que impulsa las prácticas de justicia abierta. Esta alianza se conoce con el nombre de *Free Access to Law Movement* (FALM) y promueve el acceso a la información legal, especialmente a través de Internet y de forma gratuita (2002)[10]. Esta organización tiene su

[9] Un interesante análisis sobre las leyes nacionales y provinciales de apertura de datos del sector público puede consultarse en Gómez Zavaglia (2020).
[10] Los programas en marcha de este movimiento pueden consultarse en su sitio web http://www.falm.info/

capítulo en la Argentina, cuya membresía recae en Iljusticia, un instituto que se promociona el acceso a información judiciales en el país.

Por información judicial, se entiende a todo el corpus de leyes, normativas, jurisprudencia, resoluciones y documentos producidos por los poderes públicos (FALM, 2002). De este conjunto de información, nosotros centraremos la atención en lo que hace a la apertura de las resoluciones judiciales, que es la incumbencia de la estrategia de apertura de datos del Juzgado N° 10.

En lo que refiere a la instrumentación del acceso a las resoluciones judiciales, en la Argentina, estos documentos son de acceso público. El acceso a las sentencias se halla sancionado en la Constitución Nacional y en el Pacto Internacional de Derechos Civiles y Políticos, que establece que “toda sentencia en materia penal o contenciosa será pública, excepto en los casos en que el interés del menor de edad exija lo contrario, o en las actuaciones referentes a pleitos matrimoniales o a la tutela de menores” (art. 14.1). No obstante, su publicación on-line exige nuevos resguardos, que han sido considerados en diversas normativas. En general, la normativa sobre la publicación en Internet de este material ha seguido los criterios expuestos en las Reglas de Heredia (2003). Estas reglas, son un documento redactado en 2003 que fue el fruto de una reunión de miembros de poderes judiciales de varios países de latinoamérica, académicos y especialistas en privacidad on - line. Las mismas sirven de orientación para la implementación de leyes locales y ofrecen lineamientos mínimos para la publicación en Internet de información judicial.

Las Reglas de Heredia remarcan el derecho al acceso a la información pública, tanto para garantizar la transparencia de la aplicación de la ley, como para favorecer el conocimiento del ámbito. Entre los recaudos que establecen estas reglas, señalan que las resoluciones deben ser publicadas protegiendo los datos personales de los individuos. Pero, al igual que los casos anteriores, la definición de datos personales se halla muy circunscripta a la definición clásica de PII. No obstante, al igual que las leyes de protección de datos personales, el espíritu que guía estas recomendaciones es claro: existe un derecho al acceso a la información, pero que tiene un límite concreto que es la privacidad de los involucrados en las causas (Gómez Zavaglia, 2020). Cualquier mecanismo de difusión de tal información, debe entonces reparar en este punto.

La publicación on - line de resoluciones siguiendo las recomendaciones de las Reglas de Heredia se ha adoptado en las provincias de Rio Negro[11] y Chubut[12], aunque también hay algunos Juzgados que por iniciativa propia las están aplicando, como es el caso de los Juzgados Penales, Contravencionales y de Faltas N° 10 y N° 13 de la Ciudad Autónoma de Buenos Aires. Estos dos tribunales, han llevado la iniciativa más lejos. No sólo publican los textos de las resoluciones, sino que también publican en formato tabular datos relativos a las causas que tramitan. El objetivo perseguido es garantizar la transparencia de los procesos, publicitar los actos de gobierno, pero sobre todo abrir al público, en especial a la comunidad académica, el acceso a información de interés, capaz de generar impacto positivo en la sociedad a través de investigaciones en los campos de las ciencias jurídicas y sociales. En esta línea, ponen el foco en causas vinculadas a violencia de género, tanto para concientizar al público sobre esta problemática, como para alentar la investigación en este campo.

Este ecosistema de acceso a la información jurídica se completa en la Argentina con el Plan Nacional de Apertura de Datos y el Programa de Justicia Abierta, implementado a través del Ministerio de Justicia de la Nación. A través de este último programa, el gobierno nacional recopila diversas estadísticas relativas a las causas tramitadas en todos los tribunales del país, y publica los resultados agregados en un Portal de Datos Abiertos, que puede consultarse en http://datos.jus.gob.ar/.

1.2.2. Riesgos a la privacidad en las estrategias de datos abiertos del Juzgado N° 10

Este tribunal a cargo del Dr. Pablo Casas es uno de los principales impulsores de las políticas de acceso abierto a la información judicial en la Argentina. Por un lado, con estas políticas de gobierno abierto se busca garantizar la transparencia de la justicia, y fomentar la confianza en el poder judicial. Por otro lado, se busca abrir al público el acervo de datos disponibles en las causas tramitadas por el Juzgado, con la finalidad de favorecer la investigación y el avance del conocimiento científico en diversos ámbitos. Para ello, el Juzgado se vale de herramientas on - line, a través de las cuales hace pública información general, los textos de las resoluciones, y los sets de datos antes referidos. El medio de comunicación central para acceder a toda esta información es la cuenta de Twitter del Juzgado, `@jpcyf10`[13].

[11] Implementado a través de la Acordada del Tribunal Superior de Justicia de Rio Negro N° 112 de 2003.
[12] Implementado a través del Acuerdo Plenario del Tribunal Superior de Justicia de Chubut 3701 de 2008.
[13] Puede accederse a la cuenta a través del siguiente link: https://twitter.com/jpcyf10

Como hemos comentado anteriormente, el acceso público a los textos de las resoluciones es un requerimiento por ley. No obstante, el Juzgado va un paso más allá de este mandato y, siguiendo las sugerencias de las Reglas de Heredia, publica las resoluciones on - line utilizando mecanismos básicos de anonimización para proteger la privacidad de los involucrados. Al respecto, debemos remarcar que si bien el hecho de disponer el texto on - line facilita el acceso a la información, esta siempre revistió carácter público. La pregunta relevante en este caso, teniendo en cuenta las nociones de privacidad que hemos presentado en el apartado anterior, es si la publicación on - line de este material, constituye una violación a las expectativas de circulación de la información de parte de los involucrados. De forma provisoria nuestra respuesta es que no. Aunque esta información es susceptible de ataques como los expuestos en la sección anterior, creemos que los individuos no tienen una expectativa de privacidad tan estricta desde el momento en que por ley, las resoluciones son de acceso público. El hecho de publicar el texto en Internet facilita el acceso de parte de terceros, pero no creemos que los riesgos que emanan de esta publicación sean mayores que el beneficio social que se obtiene de su publicación. La alternativa, que es la no publicación on - line del material, tampoco está exenta de riesgos. Ante requerimiento, los tribunales deberían entregar copias de las resoluciones, con lo que, en caso de un ataque de algún tipo, la diferencia entre disponer el material on - line o tener que solicitarlo (por ejemplo vía correo electrónico), no es muy considerable. De todos modos, reconocemos que este es un punto problemático, que puede ser re pensado en el futuro.

De forma complementaria a la publicación de los textos de las resoluciones, el Juzgado publica un dataset de formato tabular con información relativa a las causas, conteniendo datos relativos a los acusados, a las víctimas, a características de los delitos, a las resoluciones y ciertos datos administrativos y del proceso. Para un análisis más detallado de este dataset, remitimos al lector a la entrada 2.2 de la tesis, donde se realiza un análisis pormenorizado del dataset, y al anexo I de la tesis, donde se encuentra documentación relativa al dataset producida por el propio Juzgado.

Lo que es importante en este punto, es que este tipo de publicación de información, si bien no incluye ningún PII, representa severos riesgos a la privacidad de los involucrados. Como hemos repasado en el apartado anterior, este tipo de información es especialmente susceptible a los ataques de identificación. Esto es claramente una amenaza a la privacidad de los individuos, ya que si bien la publicación del texto de las resoluciones podría haber estado dentro de las expectativas razonables de los involucrados, la publicación de un dataset tabular online con información de las causas, claramente no cabe dentro de lo que

es una expectativa razonable. Es decir, los individuos involucrados en las causas, no tienen intención ni consienten este uso de la información, mucho menos cuando no es una exigencia legal. En este caso, a pesar de la clara ventaja para el público de acceder a estos conjuntos de datos, los riesgos a la privacidad superan con creces estos beneficios.

Es aquí donde recurriremos a ***Differential Privacy***, un enfoque que permite superar algunos de estos inconvenientes, permitiendo distribuir información potencialmente sensible, sin representar grandes riesgos a la privacidad de los individuos. Dado que la publicación del set de datos tal como la lleva a cabo el juzgado pone en riesgo la privacidad de los invididuos, nosotros nos valdremos de los aportes de ***Differential Privacy,*** para presentar una estrategia que permita la publicación de esta información, pero sin incurrir en riesgos a la privacidad de los involucrados.

1.3. *Differential Privacy*

1.3.1. Una técnica de protección de la privacidad: *Differential Privacy*

Differential Privacy (DP en adelante), no se refiere a ningún algoritmo, mecanismo ni implementación en particular. Por el contrario, DP es una **garantía**. Como tal, diversos algoritmos o mecanismos pueden ajustarse a dicha **garantía**. El fundamento de DP, son las garantías, las restricciones, y las promesas que emanan de su definición formal. Luego, todo desarrollo que se ajuste a dichas **restricciones**, puede considerarse un mecanismo de DP (Dwork & Roth, 2014).

La promesa de DP es que la inclusión o exclusión de un individuo en una base de datos, no revelará ninguna información (más allá de cierto umbral cuantificable) sobre el mismo. **Esto significa que un individuo puede confiar en que una base de datos, a través de las restricciones impuestas por DP, no revelará ningún dato personal, privado o sensible**. En este sentido, la principal ventaja de DP sobre otros mecanismos de protección de la privacidad, es que ofrece una garantía formal y verificable acerca de la cantidad de información personal que expone la base de datos (Dwork & Roth, 2014).

Uno de los aspectos interesantes de la garantía de DP, es que la posibilidad de ataques contra la privacidad no depende en forma alguna del conocimiento del oponente, de otra información disponible actualmente o en el futuro, ni de potenciales desarrollos computacionales que faciliten nuevos y más potentes ataques[14]. Por estos motivos, las técnicas de DP han ganado en popularidad desde el trabajo seminal de 2006 de Dwork *et. al.*.

Dada la calidad de las garantías de DP y la flexibilidad que permite su implementación, este principio ha sido adoptado para muchos desarrollos. Aunque inicialmente DP nació para garantizar la privacidad en entornos de bases de datos relacionales, su utilización fue extendiéndose a otros campos. Actualmente, DP se utiliza en diversos campos, como la publicación de datos, por ejemplo en iniciativas de datos abiertos o incluso en el censo 2020 de los Estados Unidos, la recopilación privada de datos, como la que realiza Google Chrome, el machine learning privado e incluso la generación de set de datos artificiales. Naturalmente cada dominio tiene sus particularidades y desafíos, y las estrategias de

[14] Esta es una diferencia con las técnicas de criptografía, cuya garantía es que en ningún tiempo aceptable un oponente puede descifrar las claves para acceder a la información.

implementación son muy diferentes, no obstante, todas ellas logran cumplir con las garantías formales de DP.

1.3.2. Definiendo *Differential Privacy*

Según hemos señalado anteriormente, **DP es una garantía**, no un algoritmo específico. Lo que decimos, es que un algoritmo o mecanismo cumple **con** las restricciones de DP. Un enfoque intuitivo a la definición de DP nos propone lo siguiente.

> Siendo D una base de datos original con n registros y D' una base de datos que contiene n-1 registros de la original, decimos que un mecanismo-función-consulta cumple con DP si:
>
> $$f(D) \cong f(D')$$

Esta definición intuitiva nos dice que un mecanismo cumple con DP si al correrse sobre una base de datos que contiene todos los registros, su resultado no difiere mucho respecto a si se hubiese corrido sobre una base de datos que contiene los mismos registros, excepto los de un individuo. Si la salida del mecanismo sobre ambas bases de datos es bastante parecida, entonces no estará revelando prácticamente información sobre ese individuo de diferencia.

Para comprender las implicancias básicas de este enunciado, pongámonos en el lugar de un atacante que desea conocer la composición de la base de datos. Supongamos que el atacante sospecha que un individuo fue incluido en cierta base de datos. Para hacer el caso más real, situémoslo en el contexto de nuestro trabajo. Imaginemos que un atacante sospecha que un vecino ha cometido un delito y que probablemente ese delito haya sido Juzgado en el tribunal N° 10 de CABA. En ese caso, el atacante tiene cierta sospecha de que su vecino va a estar incluido en el conjunto de datos que publica el Juzgado. Si este conjunto de datos fuese publicado sin ningún tipo de procesamiento, el atacante muy rápidamente podría identificar a su vecino entre los registros. Pero supongamos que el dataset ha sido publicado utilizando técnicas de DP. En ese caso, luego de la publicación del dataset, el atacante no puede decidirse si su vecino fue incluido o no. Porque si bien él tenía una sospecha inicial de que su vecino iba a estar incluido, luego de revisar efectivamente el conjunto de datos (de forma posterior a la aplicación de DP), no puede terminar de definirse. Es decir, la salida del mecanismo de DP no le permitió ganar información alguna sobre su vecino, al menos, más allá de cierto umbral. Conocer la salida del mecanismo de DP, no le permite a nuestro atacante discernir si el mecanismo fue corrido sobre D (la base de datos original, que efectivamente **sí** incluye a su vecino) o sobre D' (una base de datos que **no**

incluye a su vecino). Y el motivo por el que no puede decidirse, es porque ambas salidas se parecen mucho. De esta manera, el mecanismo de DP protege la privacidad de cada individuo, ya que no existe forma alguna de que el atacante gane conocimiento sobre su inclusión o no en la base de datos (más allá de cierto umbral definido).

Habiendo analizado esta definición intuitiva, pasemos a la definición formal de DP, tal como es propuesta por Dwork & Roth (2014).

> Un algoritmo aleatorio M con dominio $\mathbb{N}^{|x|}$ es e- *differentially private* para todas las salidas $S \subseteq Rango(M)$ y $x, y \in \mathbb{N}^{|x|}$ $tal\ que\ ||x-y||_1 \leq 1$:
>
> $$\Pr[M(x) \in S] \leq e^{\varepsilon} \Pr[M(y) \in S]$$
>
> Donde:
>
> M: es un algoritmo aleatorio.
> S: todas las salidas posibles del algoritmo M.
> x : todas las entradas en la base de datos original.
> y: todas las entradas de la base de datos paralela (con n - 1 entradas).
> ε: presupuesto de privacidad. Regula cuánta información dejamos filtrar.

Según esta definición, para todo par de bases de datos adyacentes x e y, y para todas las salidas posibles S, un atacante no puede distinguir cuál es la base de datos original sólo por observar los resultados, más allá de un ratio de probabilidades acotado por e^{ε}. Es decir, dado cualquier resultado, la probabilidad de que el mismo haya sido obtenido de correr M sobre x o sobre y, está limitada a un ratio de e^{ε} (Dwork & Roth, 2014).

Uno de los aspectos que hacen interesante al concepto de DP, es que podemos cuantificar la ganancia de información de un adversario. En base a esta formalización, sabemos que el conocimiento que el atacante puede ganar de conocer el resultado S está limitado a este ratio al cual nos referimos anteriormente. Por otro lado, la rigidez de esta garantía puede ser controlada por el parámetro ε, con lo que podemos ajustar el nivel de privacidad acorde a las necesidades del dominio. Un mecanismo de DP, puede garantizar niveles altos de privacidad, como niveles muy bajos, acorde al presupuesto definido. Más allá del nivel de privacidad utilizado, el gran aporte que está haciendo esta definición es que podemos ofrecer una garantía de que, sea la definición de privacidad rígida o laxa, no se va a estar filtrando más información que la que el curador haya decidido previamente (Dwork & Roth, 2014).

1.3.3. La sensibilidad de un mecanismo de *Differential Privacy*

Teniendo en claro cuales son las restricciones y garantías que ofrece DP ¿Cómo se implementa un algoritmo que cumpla con tales requisitos? Como dijimos anteriormente, existen diversas alternativas para implementar un mecanismo de DP, básicamente de acuerdo a la naturaleza de los datos que vayamos a trabajar y al tipo de uso que queremos darle a la información. Pero todos ellos parten de un concepto fundamental. Un mecanismo de DP va a distorsionar la base de datos original mediante la adición de ruido, en una escala suficiente como para 'enmascarar' la inclusión de un registro individual en dicha base de datos.

> La implementación básica de un mecanismo de DP M sobre una base de datos db sigue la siguiente forma:
>
> $$M = Consulta(db) + ruido$$

Supongamos que tenemos una base de datos db conteniendo información relativa a zoológicos en la ciudad y queremos realizar una consulta para saber cuántos animales existen en el zoológico más cercano. Dicha consulta sobre db nos indica que la cantidad de animales en el zoológico más cercano es de 47. Ese es el valor real. Pero el mecanismo de DP no va a devolver este valor tal cual, sino que va a adicionarle ruido. Por ejemplo, la salida del mecanismo de DP para esta consulta nos puede decir que en el zoológico más cercano hay 48 animales. Al agregarle ruido, 'enmascara' la inclusión de determinado animal en db. Imaginemos que el atacante sabía que la semana pasada el zoológico tenía 47 animales. Con el resultado actual de 48, el atacante no tiene forma de saber si el zoológico incorporó un nuevo ejemplar o no. Mejor dicho, esta salida le llevará a suponer que el zoológico **sí** incorporó un nuevo ejemplar. Pero si la salida hubiese sido 46, ese resultado lo hubiese llevado a creer que el zoológico trasladó a uno de sus animales. De cualquier modo, este atacante no podrá discernir fehacientemente sobre la existencia o no de este animal en el zoológico, y probablemente tomará una decisión errada sobre el número real de animales.

Resulta evidente que en la implementación de un mecanismo de DP la escala del ruido a adicionar es un parámetro de primer orden a considerar. Si lo que se busca es 'enmascarar' la inclusión de una observación en el registro, la escala del ruido debe guardar alguna relación con el delta que dicha observación puede tener sobre la consulta. Conceptualmente este ruido debe ser igual o mayor al impacto que la inclusión o remoción de un individuo puede tener sobre el resultado de la consulta realizada. Es aquí donde entra el concepto de sensibilidad.

Para el ejemplo del zoológico, la sensibilidad de la consulta es de 1. Esto es porque el impacto máximo que puede tener sobre el conteo la inclusión o remoción de un animal, es 1. Pero veamos el siguiente ejemplo para terminar de construir la definición de sensibilidad. Tenemos el siguiente set de datos sobre hábitos alimenticios de las personas. El mismo se compone de tres variables categóricas, cada una con dos niveles, que serán representadas utilizando one-hot-encoding.

Tabla I. Hábitos alimenticios de los individuos

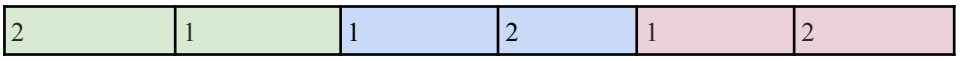
Nombre	Edad (en años)		Preferencia plato		Comidas Diarias	
	<= 40	> 40	Pastas	Carne	Tres	Cinco
Carlos	1	0	0	1	0	1
Martina	0	1	0	1	0	1
Lautaro	1	0	1	0	1	0

En base a este conjunto de datos, consideremos una consulta que consiste en un conteo de individuos que tienen más de 40 años. Rápidamente, llegamos a la conclusión de que la inclusión o remoción de cualquiera de los tres individuos de la tabla, pueden afectar en máximo en una unidad el conteo. Ahora, si en lugar de esta consulta hiciésemos una serie de consultas que nos devolvieran el conteo de observaciones para todas las columnas ¿Cuál sería la sensibilidad? ¿En cuánto puede variar el resultado de este conjunto de consultas si removemos o incluimos un individuo? En este caso, la sensibilidad sería de 3, porqué de las 6 columnas que tenemos, un individuo como máximo afectar 3 de ellas. O bien este individuo tiene más o menos de 40 años. O bien prefiere pasta o carne. O bien hace tres o cinco comidas diarias.

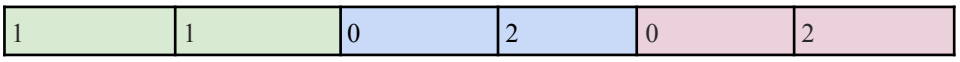
Analizando con mayor detalle este ejemplo, vemos que esta ‘gran’ consulta que realizamos previamente son en realidad 6 consultas independientes entre sí. En realidad, lo que hicimos fue realizar 6 consultas, 6 conteos, uno sobre cada columna de la tabla. El punto relevante a considerar, es que no siempre la sensibilidad de un conjunto de consultas de conteo va a ser equivalente a la cantidad de consultas individuales que se realice, sino a la cantidad de **consultas correlacionadas** que se realicen. Este nuevo concepto es importante para seguir construyendo la definición de sensibilidad. Cuando se considere un grupo de consultas, para la determinación de la sensibilidad se deben tener en cuenta los conjuntos de consultas que estén correlacionadas entre sí. O dicho de otro modo, en cuántas consultas a la vez puede

influir la inclusión o remoción de un individuo. Todas aquellas consultas que sean afectadas por la inclusión o remoción de un individuo, serán consultas correlacionadas.

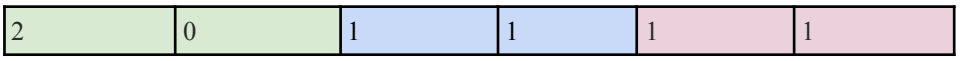
Antes de avanzar en el desarrollo del concepto de sensibilidad, veamos en detalle lo que pasa en este conjunto de consultas al ir excluyendo a cada uno de los individuos del dataset. Primero consideremos el conjunto de consultas ejecutadas sobre el dataset original. El vector conteniendo el conteo de cada columna sería este:



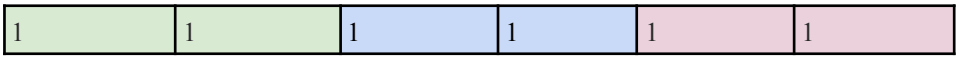
Si consideramos a Carlos y Martina, y dejamos de lado a Lautaro, el vector resultante sería este.



Y si en lugar de excluir a Lautaro, excluimos a Martina, el vector resultante del conteo se vería así:



Haciendo lo mismo, pero reponiendo a Martina y excluyendo a Carlos, el vector quedaría así:



Lo que acabamos de hacer es considerar el conjunto de consultas evaluadas en x, la base de datos original, y en todas las y posibles, es decir, en todas las bases de datos adyacentes posibles conteniendo n-1 registros de la original. Habiendo evaluado las consultas en todas estas bases de datos paralelas, podemos pasar a estimar la sensibilidad de la función.

Lo que hasta ahora estuvimos mencionando como 'la diferencia en el conteo' es en realidad la máxima norma L1 de la diferencia entre las salidas de las consultas realizadas sobre x y todas las y posibles. Según esta definición, la sensibilidad de un conjunto de funciones M es la siguiente:

$$Sensibilidad_{L1} = \max ||M(x) - M(y)||_1$$

En la práctica, la representación de las salidas de forma vectorial puede servirnos para evaluar esta definición de sensibilidad cuando tenemos un conjunto de consultas y no una única consulta. En ese caso, definimos sensibilidad como la máxima norma L1 de las diferencias de las sumatorias de los vectores resultantes de evaluar M sobre x y sobre todas las y posibles.

$$Sensibilidad_{L1} = \max \left|\left| \sum_{i=1}^{n} M(x)_i - \sum_{i=1}^{n} M(y_i) \right|\right|_1$$

En base a esta definición, concluimos (como lo habíamos hecho antes de forma más intuitiva) que para el ejemplo de más arriba, la máxima norma L1 de las diferencias entre los 4 vectores que representan todas las salidas posibles de M, es 3.

Luego, tenemos otra definición de sensibilidad, que en lugar de utilizar la norma L1, utiliza la norma L2 de la máxima diferencia entre todas las salidas del mecanismo evaluado en x y todas las y posibles. La definición de sensibilidad basada en la norma L2 es básicamente la misma que la anterior. Para una única consulta, tenemos lo siguiente:

$$Sensibilidad_{L2} = \max ||M(x) - M(y)||_2$$

Y para un conjunto de consultas:

$$Sensibilidad_{L2} = \max \left|\left| \sum_{i=1}^{n} M(x)_i - \sum_{i=1}^{n} M(y_i) \right|\right|_2$$

Utilizando esta segunda definición de sensibilidad, para nuestro ejemplo anterior en lugar de un valor de 3, obtendríamos un valor de $\sqrt{3}$. Como es de esperar, mientras que el orden de progresión de la sensibilidad medida a través de la norma L1 para k consultas correlacionadas es también k, medida a través de la norma L2 es de $\sqrt{k}$. No obstante, la elección de una u otra medida de sensibilidad no es un hecho fortuito. Si bien todavía no incursionamos en el análisis de los mecanismos de DP, la realidad es que algunos mecanismos requieren el uso de una medida de sensibilidad basada en la norma L1, como

el laplaciano, y otros, como el gaussiano, pueden garantizar DP utilizando una definición de sensibilidad más acotada, basada en la norma L2. Aunque no entraremos en el desarrollo de los cálculos que prueban esto, la idea de fondo es que utilizando el mecanismo laplaciano, el nivel de ruido que debe incorporarse, para cumplir con inecuación que define DP, debe estar en función de la norma L1 de la sensibilidad. En cambio, utilizando el mecanismo gaussiano, pueden satisfacerse estas mismas restricciones utilizando la sensibilidad basada en la norma L2 (Dwork & Roth, 2014).

De este conjunto de definiciones se desprenden las siguientes conclusiones, de especial relevancia para la fase de implementación de los mecanismos de DP propuestos en la segunda parte del trabajo. Si la salida de una función de conteo es un vector de d dimensiones, de las cuales hay un máximo de k dimensiones (consultas) correlacionadas, la sensibilidad L1 será la suma de las sensibilidades individuales de cada uno de estos k conteos. Como sabemos que la sensibilidad de un conteo es uno, esto significa que la sensibilidad de esta función de conteo será k. En cambio, si utilizamos la sensibilidad L2, para esta misma función la sensibilidad será de $\sqrt{k}$. [15]

La medición de la sensibilidad puede realizarse en dos contextos. En el ejemplo anterior donde comparamos todas las salidas posibles de M evaluado en x y en todas las y posibles, hicimos un análisis a nivel local. Es decir, medimos la sensibilidad sobre el conjunto de datos que estábamos trabajando. Este es el concepto de **sensibilidad local** (Li *et. al.*, 2017). Para la estimación, sólo se tienen en cuenta los datos con los que se está trabajando. Esto tiene dos ventajas. La primera, es que puede realizarse una medición empírica en base a los datos, como la que realizamos más arriba. En segundo lugar, es posible que la sensibilidad, medida sobre un único conjunto de datos de entre todos los posibles, sea menor que si se trabajase con un enfoque más amplio de sensibilidad. Si bien en algunas aplicaciones tiene sentido trabajar con este criterio de sensibilidad, en general no es aconsejado. La medición local de la sensibilidad tiene sentido en un escenario en particular, que es el de *Local Differential Privacy* (que será abordado en la sección 1.4.1), donde la información debe privatizarse antes de que que sea recopilada por el curador central, por ejemplo en sistemas que recolectan información de usuarios como dispositivos móviles o exploradores de Internet[16].

[15] Un análisis de este tipo de funciones puede consultarse en Fathima (2020). Un caso similar es expuesto en la sección 'L1 VS L2 Sensitivities'. Un desarrollo matemático más completo, pero en igual sentido, puede verse en Desfontaines (2021).
[16] Al respecto, recomendamos ver Fathima (2020).

El criterio de medición de la sensibilidad que generalmente se utiliza es el de **sensibilidad global** (Fathima, 2020). Esto excede la medición sobre un único conjunto actual de datos, e incluye todas las posibles diferencias entre todos los posibles datasets adyacentes conteniendo n-1 registros. Esta definición de sensibilidad es independiente del dataset concreto con el que se esté trabajando y depende únicamente de la consulta que se esté realizando. Esta definición de sensibilidad requiere de parte del curador un amplio conocimiento del dominio. Ya no puede valerse del conjunto de datos del que dispone, sino que debe anticipar cómo va a variar la consulta que realiza en función de datos que no tiene (pero cuyo dominio - en teoría al menos - conoce) . Por ejemplo, si en lugar de un simple conteo sobre una variable categórica la consulta fuese del tipo ¿Cuál es el sueldo más alto entre los asalariados de CABA? La determinación de la sensibilidad sería más difícil de definir. Dejando de lado el hecho de que no existe teóricamente un límite superior al dominio de esta variable (aunque lógicamente se puede truncar) ¿Cómo puede el curador anticipar en cuánto va a variar el resultado de la consulta al incluir o excluir un individuo? La única forma de anticiparlo es a partir de un conocimiento profundo del dominio (Li *et. al.*, 2017).

1.3.4. Los mecanismos de *Differential Privacy*

Un mecanismo de DP es un algoritmo capaz de responder consultas sobre un conjunto de datos ateniéndose a las restricciones de DP (Dwork & Roth, 2014). Como mencionamos previamente, cualquier implementación que cumpla con estas restricciones, puede ser considerado un mecanismo de DP. Por ello, este es un campo en permanente innovación, donde se realizan avances constantes en diversas áreas y surgen nuevos algoritmos capaces de implementar DP para distintos usos y contextos. No obstante, existen algunos mecanismos básicos que son los más usados y sientan los fundamentos de desarrollos más complejos. En este apartado consideraremos los mecanismos laplaciano y gaussiano, que son los mecanismos fundamentales de DP, aunque existen muchos más. Ambos mecanismos pueden ser utilizados para responder consultas realizadas sobre variables cuantitativas (discretas o continuas) cuya respuesta sea también numérica. Entre las consultas a las que pueden aplicarse tenemos conteos, sumas o promedios. En cambio, no pueden utilizarse para responder consultas del tipo ¿Cuál es el color de pantalones más vendido esta temporada?[17]

1.3.4.1. El mecanismo laplaciano

Recordemos la implementación básica de un mecanismo de DP que dimos más arriba:

[17] Para responder consultas de este tipo, debe utilizarse el mecanismo exponencial.

La implementación básica de un mecanismo de DP M sobre una base de datos db sigue la siguiente forma:

$$M \ = \ Consulta(db) \ + \ ruido$$

La particularidad del mecanismo laplaciano es que toma ruido de una distribución laplace y lo adiciona a la salida de la consulta. Si la salida de la consulta (o del conjunto de consultas), es un vector de d dimensiones, el mecanismo adicionará ruido a cada una de estas dimensiones de la salida (Dwork & Roth, 2014: Kamath, 2020a).

Una función que aplica el mecanismo laplaciano, puede expresarse como:

$$f \ : \ \mathbb{N}^{|x|} \ \Rightarrow \mathbb{R}^k$$

$$M_L(x, \ f(.), \ \varepsilon) \ = \ f(x) \ + \ (Y_1, \ldots, \ Y_k)$$

Donde:

Y_i son variables aleatorias I.I.D. tomadas de una distribución $Lap\ (s_1\ /\varepsilon)$
s_1 : es la sensibilidad L1 de la función.
ε : es el presupuesto de privacidad. Regula cuánta información dejamos filtrar.

Es importante destacar que este mecanismo considera para la escala de la distribución de la que toma el ruido la sensibilidad L1, a diferencia del mecanismo gaussiano que veremos a continuación, que utiliza la sensibilidad L2.

1.3.4.2. El mecanismo gaussiano

Conceptualmente, el mecanismo gaussiano se parece mucho al laplaciano. Pero, a diferencia de éste, el mecanismo gaussiano no puede garantizar salidas ε - *differentially private* sino que debe relajar un poco esta definición. La garantía que nos va a dar el mecanismo gaussiano respecto a la salida es que será $\varepsilon, \ \delta$- *differentially private.* Este nuevo parámetro δ que agrega la definición es una relajación a la garantía de privacidad y conceptualmente representa la probabilidad de que un evento adverso de filtrado de información ocurra (Dwork & Roth, 2014; Kamath, 2020b; Desfontaines, 2021). A priori, esta relajación no debiera preocuparnos demasiado, ya que valores bajos de δ no representan más que una remota chance de eventos adversos. Veamos a continuación como queda

reformulada la definición de DP que realizamos a inicios del apartado en función de esta relajación introducida.

> Un algoritmo aleatorio M con dominio $\mathbb{N}^{|x|}$ es e, ε, δ- *differentially private* para todas las salidas $S \subseteq Rango(M)$ y $x, y \in \mathbb{N}^{|x|}$ *tal que* $||x - y||_1 \leq 1$:
>
> $$\Pr[M(x) \in S] \leq e^{\varepsilon} \Pr[M(y) \in S] + \delta$$
>
> Donde:
>
> M: es un algoritmo aleatorio.
> S: todas las salidas posibles del algoritmo M.
> x : todas las entradas en la base de datos original.
> y: todas las entradas de la base de datos paralela (con n - 1 entradas).
> ε: presupuesto de privacidad. Regula cuánta información dejamos filtrar.
> δ: es la probabilidad de un evento adverso de filtrado de información.

Nótese que la única diferencia con la definición inicial de DP es que hemos introducido el parámetro δ, que regula la probabilidad de un evento adverso de filtrado de información. De ahora en más, cuando nos refiramos a mecanismos de DP, siempre tendremos en cuenta esta definición ampliada. En el caso del mecanismo laplaciano, donde no hay posibilidad de eventos adversos de filtrado de información se considera que el parámetro δ es siempre 0.

Los dominios y codominios que toma una función que implementa el mecanismo gaussiano son los siguientes:

$$f : \mathbb{N}^{|x|} \Rightarrow \mathbb{R}^k$$

En cuanto a su funcionamiento, el mecanismo gaussiano, al igual que el laplaciano, lo hace es adicionar ruido a cada punto del vector de salida de f. Pensemos provisoriamente que una función que implementa este mecanismo tiene la siguiente forma:

$$F(X) = f(x) + N(0, \sigma)$$

Básicamente, lo que dice esta última definición es que el mecanismo gaussiano computa la salida de una función f y agrega ruido a cada punto de la salida tomado de una distribución normal centrada en 0 y con cierto σ. La varianza de esta distribución va a estar dada por:

$$\sigma^2 \;=\; \frac{2\Delta_2^2 \ln\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}$$

A diferencia del mecanismo laplaciano, este utiliza en la determinación de la escala de la distribución la sensibilidad L2.

Con todos los elementos en su lugar, podemos formalizar la definición del mecanismo gaussiano (Kamath, 2020b):

$$f \;:\; \mathbb{N}^{|x|} \;\Rightarrow \mathbb{R}^k$$

$$M_G(x,\; f(.),\; \varepsilon,\; \delta) \;=\; f(x) \;+\; (Y_1,\; \ldots,\; Y_k)$$

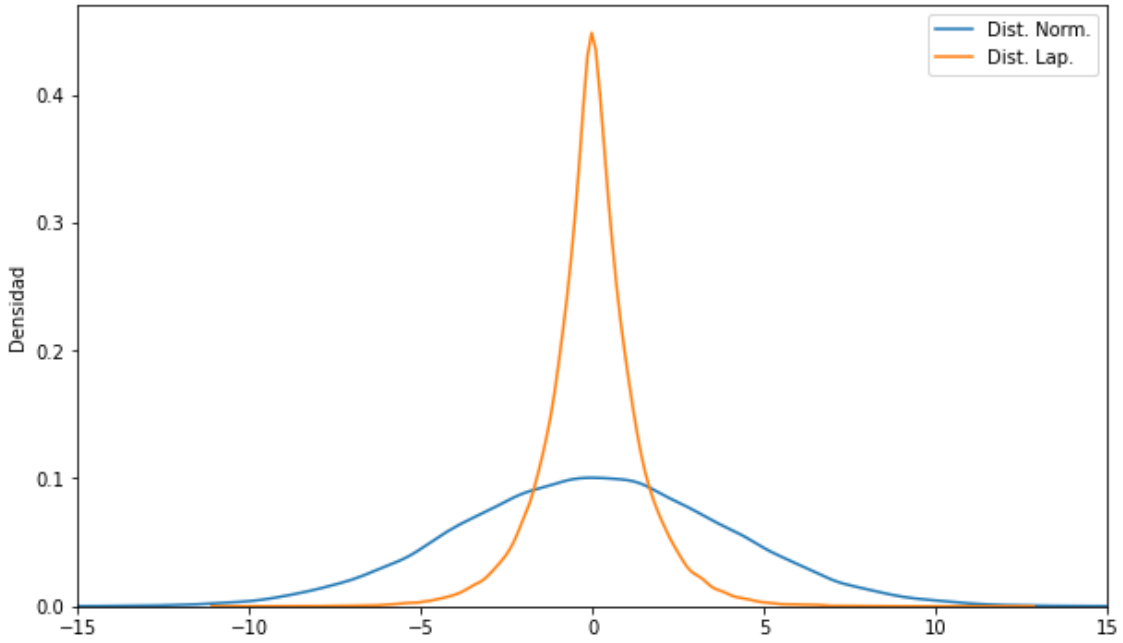
Donde Y_i son variables aleatorias I.I.D. tomadas de una $N(0,\; \sqrt{\frac{2s^2 \log\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}})$

1.3.4.3. Comparativa de los mecanismos laplaciano y gaussiano

Según hemos señalado, estos mecanismos funcionan de forma similar. No obstante, difieren en la definición de sensibilidad que utilizan y en la distribución de la que generan el ruido ¿Qué consecuencias tienen estas diferencias? Como veremos a continuación, el mecanismo laplaciano funciona mejor para consultas de baja sensibilidad, mientras que el mecanismo gaussiano tiene mejores resultados en consultas de alta sensibilidad[18].

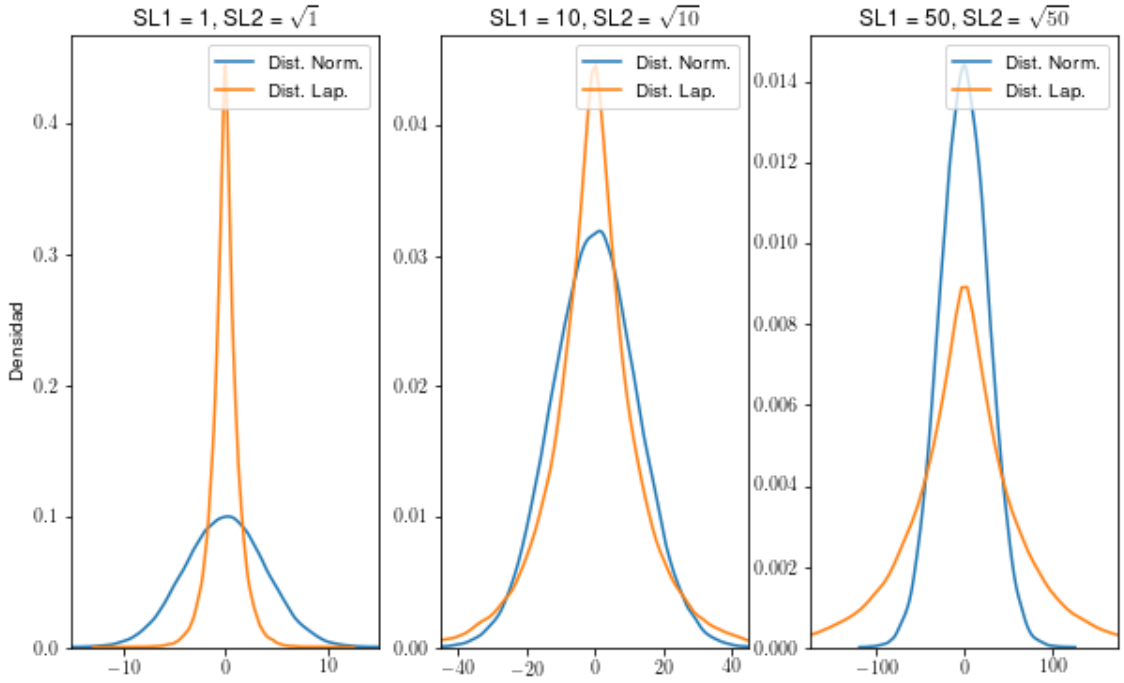
[18] Recordemos de las definiciones presentadas en 1.3.3 que la sensibilidad de una consulta o función representa el delta máximo en la salida al incorporar o eliminar el registro de un individuo en el conjunto de datos. Conceptualmente, este delta se vincula con la cantidad de ruido que hay que incorporar a las salidas privadas para poder 'enmascarar' la inclusión o no de cierto registro. El mecanismo laplaciano utiliza como definición de sensibilidad la norma L1, mientras que el gaussiano la norma L2, por lo que a medida que el delta de una función tiende a crecer, la sensibilidad que computa el mecanismo laplaciano tienen a crecer más velozmente que la del mecanismo gaussiano. Por ello, según el delta de la función funcionará mejor uno u otro mecanismo.

Gráfico I. Comparativa de las distribuciones Normal y Laplaciana para determinados mecanismos.



Si fuésemos a agregar ruido a una consulta que es un conteo, de sensibilidad (delta) 1, utilizando un presupuesto de 1ε (y un δ de 0.0005 para el mecanismo gaussiano), las distribuciones de las que tomarían ruido los mecanismos laplaciano y gaussiano serán las que vemos en el gráfico I. Se observa para este caso que la dispersión del ruido tomado de la distribución gaussiana va a ser mucho más grande que si se toma de la distribución laplace. Pero, este no siempre va a ser el caso.

Gráfico II. Comparativa de los mecanismos Gaussiano y Laplaciano bajo diferentes sensibilidades de consulta



El gráfico II muestra cómo van cambiando las distribuciones a medida que aumenta el delta de la consulta. Para el ejemplo anterior, teníamos una única consulta de sensibilidad (delta) 1. Tanto usando la norma L1, como la norma L2 en la definición de sensibilidad, esta va a ser siempre 1. Pero, a medida que crece la cantidad de consultas que queremos realizar, y aumenta el delta de la consulta, esta situación va a ir cambiando a favor de la sensibilidad L2. Por ejemplo, si realizamos 10 consultas de conteo, de sensibilidad individual de 1, la sensibilidad L1 va a ser también de 10, mientras que la sensibilidad L2 va a ser de $\sqrt{10}$. Pero esta brecha se hace más notoria si incrementamos la cantidad de consultas. Realizando 50 consultas de sensibilidad (delta) individual 1, tenemos que bajo el cómputo de sensibilidad L1, esta es de 50, mientras que computando la sensibilidad L2, esta es de $\sqrt{50}$. ¡Es una gran diferencia! La sensibilidad L2, es menos de un 20% de lo que es la sensibilidad L1. Como es de esperar, esto tiene implicancias en la dispersión de las distribuciones de donde los mecanismos toman el ruido.

Para consultas de bajo delta, la distribución laplace ofrece significativas ventajas. Debido a que la distribución cae rápidamente, y que la distribución normal es más 'achatada', el mecanismo laplaciano incorpora ruido mucho menos disperso. Como se ve en el gráfico II, con una sensibilidad de 1, la distribución laplaciana tiene una densidad acumulada cerca del centro muy superior a la normal. De este modo, puede garantizar DP adicionando menos ruido a las consultas.

Para una sensibilidad L1 de 10, y una sensibilidad L2 de $\sqrt{10}$, las distribuciones laplace y normal se asemejan bastante. Es de esperar que el ruido incorporado por ambos mecanismos sea aproximadamente similar. No obstante, ya vemos una característica de la laplace relevante: sus colas anchas. Ya en este ejemplo, estamos viendo que si bien el centro de las distribuciones se parece, en los extremos las colas de la laplaciana están acumulando más densidad que la normal. Esto implica que esta distribución tiene mayores probabilidades de tomar ruido de los extremos.

Las colas anchas de la distribución laplaciana resultan evidentes en el último ejemplo del gráfico II. Para el caso de 50 consultas de sensibilidad individual de 1, tenemos una sensibilidad L1 de 50 y una sensibilidad L2 de $\sqrt{50}$. En este caso, la normal acumula el grueso de la densidad cerca del centro, mientras que las colas largas de la laplaciana acumulan una proporción importante de la densidad en los extremos. En la práctica esto significa que la dispersión del ruido será mucho mayor si se utiliza el mecanismo laplaciano que el gaussiano.

A pesar de las ventajas del mecanismo gaussiano, no debemos de perder de vista que el mismo no puede garantizar DP pura, ya que siempre incorpora el parámetro δ que representa una probabilidad de eventos de goteo de información. No obstante, la conclusión general es que siempre que estemos en un contexto de sensibilidad alta, el mecanismo gaussiano dará mejores resultados, mientras que en contextos de sensibilidad baja, el laplaciano funcionará mejor. Así y todo, existe un espacio intermedio donde no es del todo claro cuál de los dos ofrecerá mejores resultados, con lo que siempre es buena práctica evaluar ambos mecanismos antes de decidirse por uno u otro.

1.3.5. Registros Correlacionados y *Group Privacy*

En el contexto de DP, dos registros correlacionados son aquellos que se hallan vinculados por algún motivo. Por ejemplo, si en un dataset sobre enfermedades infecciosas hay registros que corresponden a personas convivientes, estos registros van a estar correlacionados. Es probable que si un individuo del hogar enferme, también lo haga algún otro conviviente. Estos individuos forman parte del mismo grupo. Pero también puede pasar que un individuo aparezca muchas veces en el mismo dataset, formando también un grupo, pero de n registros vinculados a sí mismos. Este puede ser el caso de un dataset de transacciones en un e-commerce. Allí, un individuo puede haber realizado múltiples pedidos. En ese caso, todos estos registros están correlacionados. Conceptualmente, los registros

correlacionados son todos aquellos que pueden verse afectados (o impactados) por la exclusión (inclusión) de un individuo en el dataset, ya que la misma, hará variar también la salida de una consulta, y consecuentemente revelará información sobre todo el grupo de registros correlacionados. En el caso del registro de enfermedades, un único individuo enfermo puede afectar a todos sus convivientes. En el caso de un registro de transacciones comerciales, un único individuo puede realizar más de una transacción. Así, todos estos registros correlacionados revelan algo de información sobre este único individuo.

Uno de los supuestos de DP es que los registros sobre los que se trabaja no se hallan correlacionados. En el caso en que existan registros correlacionados, debe aplicarse *Group Privacy*, que básicamente lo que hace es agregar ruido en proporción al tamaño del grupo, de forma tal que no se pueda discernir acerca de la inclusión o no del grupo en el dataset. Nótese que es una extensión conceptual de la definición de DP, que tiene un enfoque similar, pero a nivel individual. El problema de esta solución es que al incrementar el ruido a agregar en el mecanismo de DP, la calidad de la solución se degrada rápidamente. Por otro lado, para poder aplicar *Group Privacy* el número máximo de registros correlacionados debe tener una cota superior. Debe existir un límite a la cantidad de registros que un sólo individuo pueda influir. Si no lo existe dentro del dominio, debe establecerse arbitrariamente, lo que supone siempre el descarte de cierta cantidad de registros en caso de que algún grupo supere la cota máxima de registros establecida para utilizar *Group Privacy*.

1.3.6. Composición

En las secciones anteriores hemos visto que un parámetro fundamental de los mecanismos de DP es el presupuesto ε, que controla la cantidad de información que un mecanismo deja filtrar. Aunque no hemos reparado en el asunto, el lector probablemente haya reparado en que el mismo tiene una dinámica acumulativa en función de la cantidad k de consultas correlacionadas que se realicen. Al fin y al cabo, si cada consulta que representa la salida de un mecanismos de DP revela una porción de información sobre el dataset, el conjunto de información que fue revelada como resultado de realizar varias consultas debe computarse de forma agregada, de alguna manera. Esto es lo que se conoce como la composición (Dwork & Roth, 2014).

1.3.6.1. Consultas secuenciales y paralelas

Ahora bien. ¿Qué consultas debemos considerar para realizar la composición del presupuesto? El primer paso para determinar el cómputo del presupuesto, es la identificación de las consultas que deberán incluirse en su estimación. En este sentido, vamos a considerar en el cálculo del presupuesto, la cantidad máxima de consultas k correlacionadas. Recordemos para el caso del conteo que realizamos sobre la tabla I, que teníamos un conjunto de 6 consultas de conteo, pero sólo habían 3 correlacionadas, ya que un individuo puede influir a la vez, como máximo, en 3 de esos 6 conteos. En este caso, debemos agregar los presupuestos de estas 3 consultas, no de 6.

Notemos que estas k consultas que estamos realizando están llegando en forma **secuencial**. Es decir, se realiza una tras la otra y siempre sobre el mismo conjunto de registros (Li *et. al.*, 2017). Consultas de este tipo siempre deben agregarse. Volviendo a nuestro ejemplo de la tabla I, si luego de finalizado el conteo queremos realizar más consultas sobre estos datos, como por ejemplo cuántos individuos de más de 40 años comen cinco veces al día, esta sería otra consulta que se respondería utilizando el mismo conjunto de datos. Esta última consulta también debería computarse en el presupuesto utilizado. En este sentido, deben considerarse todas las consultas que se respondan utilizando el mismo conjunto de datos.

Ahora imaginemos el caso de consultas que se respondan utilizando subconjuntos separados de los datos. Por ejemplo, si sobre la tabla I tenemos estas dos consultas:

- ¿Cuántos individuos de **más** de 40 años comen cinco veces al día?
- ¿Cuántos individuos de **menos** de 40 años comen cinco veces al día?

Dado que DP busca evitar la exposición de datos sobre individuos, no sobre un dataset en particular, y dado que un individuo sólo puede afectar el resultado de una de estas dos consultas (o tiene más o menos de 40 años), estos presupuestos **no** deben agregarse. Este tipo de consultas se denominan consultas **paralelas** y su presupuesto va a ser el **mayor** utilizado en cualquiera de ellas (Li *et. al.*, 2017).

1.3.6.2. Composición simple

La forma más sencilla de considerar el presupuesto de un conjunto de consultas es la composición simple. Bajo este enfoque, el presupuesto total de un conjunto de consultas correlacionadas, no es más que la agregación lineal del presupuesto individual de cada una de ellas (Dwork & Roth, 2014). Si se realizan k consultas, siendo cada una de ellas (ε, δ) *differentially - private*, tenemos que el presupuesto total está dado por:

$$\varepsilon' = \sum_{i=1}^{k} \varepsilon_i \qquad \delta' = \sum_{i=1}^{k} \delta_i$$

Donde:

ε': representa el presupuesto total utilizado.

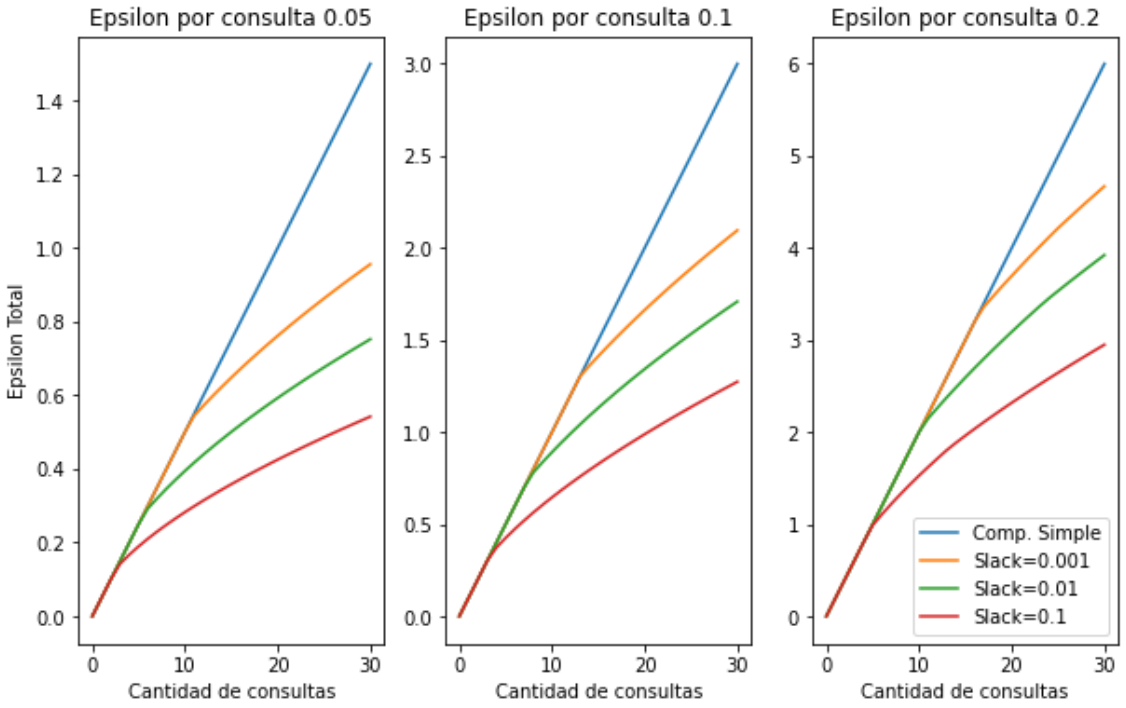
δ' : representa el delta total acumulado, es la probabilidad de un evento de filtración de datos.

k: representa el máximo número de consultas correlacionadas.

1.3.6.3. Composición avanzada

A diferencia de la composición simple, la composición avanzada, como su nombre lo anticipa, es mucho más compleja. En lugar de realizar una agregación lineal del presupuesto, el orden de agregación es siempre menor o igual a k. Esto permite que a medida que va creciendo la cantidad de consultas a realizar, el presupuesto total sea bastante inferior que utilizando el teorema de composición simple. Por supuesto, esta ventaja no viene exenta de un 'costo'. Calcular el presupuesto utilizando el teorema de la composición avanzada requiere una relajación extra de la privacidad, ya que incurre en una probabilidad adicional de evento de goteo de datos. Esta probabilidad de goteo de datos se denomina slack, y lo representaremos con δ''. Este slack, va a adicionarse al δ' que teníamos previamente. Entonces, si bien el teorema de composición avanzada otorga ciertas ventajas, estas incurren en un costo extra, que hay que saber balancear para poder mantenerse dentro de un rango de privacidad aceptable. Al igual que a la hora de seleccionar un mecanismo de DP, la conveniencia o no de computar el presupuesto como composición simple o avanzada debe realizarse considerando el caso de uso, las necesidades de privacidad, las consultas a realizar y el conjunto de datos disponible.

Gráfico III. Comparativa de composición simple y avanzada utilizando diferentes epsilon y slack



El gráfico III nos permite analizar las implicancias de la composición avanzada. El mismo, muestra la comparativa del cómputo del ε' para diferentes valores de ε (presupuesto utilizado en cada consulta individual) y diferentes valores de slack (probabilidad de goteo de datos), según la cantidad de consultas realizadas. La primera conclusión que se obtiene es que a mayor cantidad de consultas, mayor es la ventaja que ofrece la composición avanzada. Asimismo, para pocas consultas, no ofrece ventajas respecto a la composición simple. Por otro lado, también vemos que la brecha entre la composición simple y avanzada tiende a disminuir a medida que se incrementa el presupuesto de cada consulta individual. Por último, vemos que a medida que incrementamos el slack, el cómputo total del presupuesto va disminuyendo (aunque lógicamente aumenta la probabilidad de goteo de datos representada por el slack).

Con estos elementos en su lugar, ya podemos presentar la definición formal del teorema de composición avanzada (Dwork & Roth, 2014). Si se realizan k consultas, siendo cada una de ellas (ε, δ) *differentially - private*, tenemos que bajo el teorema de la composición avanzada, el presupuesto total está dado por:

$$\varepsilon' = \sqrt{2k\ln(1/\delta'')\epsilon} + k\epsilon(e^{\varepsilon} - 1) \qquad \delta' = k\delta + \delta''$$

Donde:

ε' : representa el presupuesto total utilizado.

δ' : representa el delta total acumulado, es la probabilidad de un evento de filtración de datos.

δ'': representa el slack.

k: representa el máximo número de consultas correlacionadas.

1.3.7. Medidas de calidad de la información

Ya sabemos cómo implementar un mecanismo de DP. Pues bien ¿Cómo seleccionar entre ellos cuál funciona mejor? Responder esta pregunta nos lleva a preguntarnos ¿Qué es funcionar 'mejor'? Sobre este último punto, la respuesta no es difícil. La mejor implementación de DP será aquella, que dado cierto presupuesto, distorsiona menos la distribución original de los datos. O lo que es lo mismo, la que puede lograr un nivel de calidad dado, a menor presupuesto de privacidad.

Conceptualmente, tenemos dos caminos para medir la calidad de la información. El primero, pasa por medir el ruido incorporado durante la utilización del mecanismo. Así, podemos sumar el ruido aplicado en cada una de las consultas individuales y luego comparar entre mecanismos cuál es el que incorporó la menor cantidad de ruido . Otra alternativa para evaluar la calidad del modelo pasa por medir la distancia entre una salida no privada de una consulta y la salida privada. Por ejemplo, la calidad de una consulta puede medirse como $\left|f(x) - \hat{f}(x)\right|$. A menor distancia, mayor es la calidad del mecanismo (Li *et. al*., 2017). Siguiendo este abordaje, son varias las métricas que pueden utilizarse para medir esta distancia.

Una de las métricas más utilizadas para medir la distancia entre una salida no privada y una privada es el error absoluto medio. Esta medida resulta útil para la construcción de tablas de distribución de frecuencias, ya que otorga un indicador de fácil interpretación de la distancia que separa la distribución original de la nueva. El error absoluto medio (EAM), entre un vector verdadero Y y uno predicho $\hat{Y}$, se define como:

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i|$$

Otra medida utilizada es el error cuadrático medio. Esta medida es útil para comprar modelos entre sí, pero no otorga un indicador tan sencillo de interpretar, como sí lo es el error absoluto medio. El error cuadrático medio (ECM), entre un vector verdadero Y y uno predicho $\hat{Y}$, se define como:

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Otra medida que suele utilizarse es la divergencia KL. Ésta, mide la divergencia entre dos distribuciones de probabilidades. La divergencia KL entre un vector de probabilidades Y y uno predicho $\hat{Y}$, se define como:

$$D_{KL}(Y||\hat{Y}) = \sum_{i}^{n} Y_i \ln \frac{Y_i}{\hat{Y}_i}$$

Esta medida informa la cantidad de bits o nets (de acuerdo a la base del logaritmo que se utilice) necesarios para que la distribución de probabilidades $\hat{Y}$ sea igual a Y. Lo que es interesante para nosotros, es que la divergencia KL no sólo refleja la diferencia en cada punto del vector de probabilidades, sino que también permite entender si la distribución de las probabilidades son en algún rango del dominio especialmente disímiles. Por ejemplo, dos distribuciones podrían tener EAM y ECM muy bajos, pero si este error se concentra en un rango acotado del dominio de la distribución, la divergencia KL penalizará esta situación. Esta característica resulta de utilidad, ya que no sólo nos va a interesar medir el error promedio entre pares de puntos de Y e $\hat{Y}$, sino también mantener la estructura de la distribución a lo largo del dominio.

A lo largo del trabajo utilizaremos el EAM y la divergencia KL como medidas de calidad de la información. Ambas medidas se computarán entre las distribuciones originales, previa aplicación de DP, y las salidas privadas, luego de la aplicación de los mecanismos de DP. Utilizaremos el EAM principalmente para tener una medida de fácil interpretación de la 'distancia' que separa las salidas privadas de las originales. En cambio, utilizaremos la

divergencia KL para evaluar diferentes estrategias de aplicación de DP y compararlas entre sí. En este sentido, utilizaremos la divergencia KL como un criterio para evaluar la ganancia de información a medida que refinamos la implementación de los mecanismos de DP. Aunque también podemos utilizar el EAM para esta finalidad, consideramos que la divergencia KL es capaz de captar variaciones sutiles en la calidad de la información, que el EAM puede pasar por alto.

1.3.8. Propiedades de *Differential Privacy*

En base a lo visto hasta aquí, podemos destacar las principales propiedades de DP (Dwork & Roth, 2014).

- DP promete que ningún individuo va a ver vulnerada su privacidad por goteo de datos en el presente ni en el futuro, más allá del ε aceptado.
- En este sentido, ofrece garantías contra cualquier tipo de ataques, en especial de identificación.
- DP permite calcular una medida cuantitativa de la cantidad de información privada que se está ‘fugando’.
- DP permite componer mecanismos y cuantificar la pérdida global de privacidad.
- La información procesada con garantías de DP, es inmune al post procesamiento. Es decir, una vez que se aplicó un mecanismo de DP, no existe forma de recomponer la información original, con lo cual, ningún tipo de operación posterior puede suponer riesgos adicionales a la privacidad

1.4. *Differential Privacy* en la práctica

En las secciones anteriores hemos presentado los bloques fundamentales de DP. Hemos introducido la definición formal y abordado las restricciones y garantías que DP ofrece. Luego, repasamos los mecanismos laplaciano y gaussiano y vimos cómo utilizarlos. En esta sección, veremos una serie de aplicaciones y usos que se construyen en base a estos elementos fundacionales. Tales abarcan diversos dominios como la recopilación de datos privados, machine learning, data publishing y data analysis.

1.4.1. *Differential Privacy* para la recopilación de datos

En ocasiones, una entidad centralizada quiere o necesita recopilar información de terceros. Este es el caso por ejemplo de un sitio de e-commerce que recopila información de las interacciones de sus clientes, o de un software que envía estadísticas de uso a un sistema central. Lo mismo vale para la información recopilada a través de dispositivos móviles o sensores remotos. Básicamente, este modelo aplica para cualquier flujo de información donde existan múltiples fuentes y un curador central que recopila y utiliza estos datos (Kamath, 2020c; Fathima, 2020).

En general, en los modelos previos de DP que estuvimos analizando, era el curador en poder de los datos, quien, ante una consulta aplicaba DP sobre la información y devolvía una salida. En este nuevo modelo de DP, el curador ya no es de confianza, y el mecanismo de DP se aplica en el origen de los datos. Es decir, es cada usuario/fuente/productor de datos el que protege su información antes de enviársela al curador.

En el modelo anterior, el curador conocía todos los datos en bruto, y la aplicación del mecanismo de DP estaba bajo su responsabilidad. Este modelo se conoce como '*Global Differential Privacy*' y se caracteriza por la recopilación centralizada de la información en bruto, que luego, en manos de un curador confiable, va a ser utilizada de diversos modos. No obstante, hay ocasiones donde no se confía en un curador central con acceso irrestricto a los datos y se prefiere garantizar la privacidad a nivel del origen, permitiendo que el curador sólo reciba una versión ya privada de la información. Este tipo de escenarios suele darse en contextos donde empresas proveedoras de servicios recopilan información de sus usuarios y/o dispositivos, pero éstos tienen expectativa de privacidad sobre los mismos ¿A quién le gustaría que Google recopilara todo nuestro historial de navegación? O ¿Quién no se sentiría incómodo sabiendo que el fabricante de su automóvil conoce al detalle los datos de su uso, sus recorridos, horarios y pormenores de los viajes? La misma precaución vale

respecto a muchas aplicaciones de uso cotidiano, como plataformas de movilidad o e-commerce.

En tales contextos, el usuario tiene expectativa de privacidad y desea proteger su información. Pero, por otro lado, muchas de estas aplicaciones y sistemas recopilan datos ya sea para garantizar el buen funcionamiento de los mismos, como para optimizar ciertos aspectos del servicio. Es aquí, donde entra DP, que permite a estas empresas recopilar los datos, pero brindando garantías de privacidad a los usuarios. Este enfoque de DP se conoce como '*Local Differential Privacy*', que se contrapone al anterior de '*Global Differential Privacy*'.

Bajo '*Local Differential Privacy*', el mecanismo de DP se aplica a nivel local, no centralizado, y sólo se envía al curador central información ya privada. Si bien este modelo resulta interesante, ya que parece brindar garantías de privacidad mucho más sólidas que el modelo de '*Global Differential Privacy*', ya que nadie puede ver nuestra información privada, ni siquiera el curador, en la práctica estos modelos resultan difíciles de implementar. El principal obstáculo es que la cantidad de observaciones que se requieren para garantizar información de calidad luego de aplicar DP a nivel local, es de orden cuadrático respecto a la necesaria en la implementación global. Eso limita la viabilidad de la aplicación a entornos o empresas con grandes volúmenes de datos y gran cantidad de usuarios (Kamath, 2020c).

Este enfoque ha sido adoptado en varias empresas tecnológicas. Por ejemplo, Google utiliza una implementación de Local DP denominada RAPPOR, con la que recolecta datos de navegación web y de procesos del sistema operativo (Erlingsson, Pihur & Korolova, 2014). Apple también tiene su propia implementación de Local DP llamada *Private Count Mean Sketch* (Apple, 2017) y es utilizada para recolectar información sobre el uso de emojis y de los teclados de iPhone. Microsoft también tiene sus propios desarrollos para recopilar información de los usuarios sobre el uso de aplicaciones en Windows (Ding, Kulkarni & Yekhanin, 2017).

A pesar de estos y otros desarrollos, el enfoque de Local DP no es dominante, principalmente debido a la escala de datos que requiere. Por otro lado, no todos los contextos involucran un curador no confiable, con lo que una expectativa de privacidad tan estricta no siempre es real. En lo que sigue del trabajo, nos continuaremos refiriéndonos a DP siempre en el contexto de *Global DP*. Salvo estos pocos ejemplos aquí mencionados, el grueso de los desarrollos en DP van por el camino de *Global DP*. Asimismo, en lo que hace

a los objetivos de este trabajo, nos vamos a estar moviendo también en un contexto de Global DP, donde el Juzgado actúa como curador central de la información.

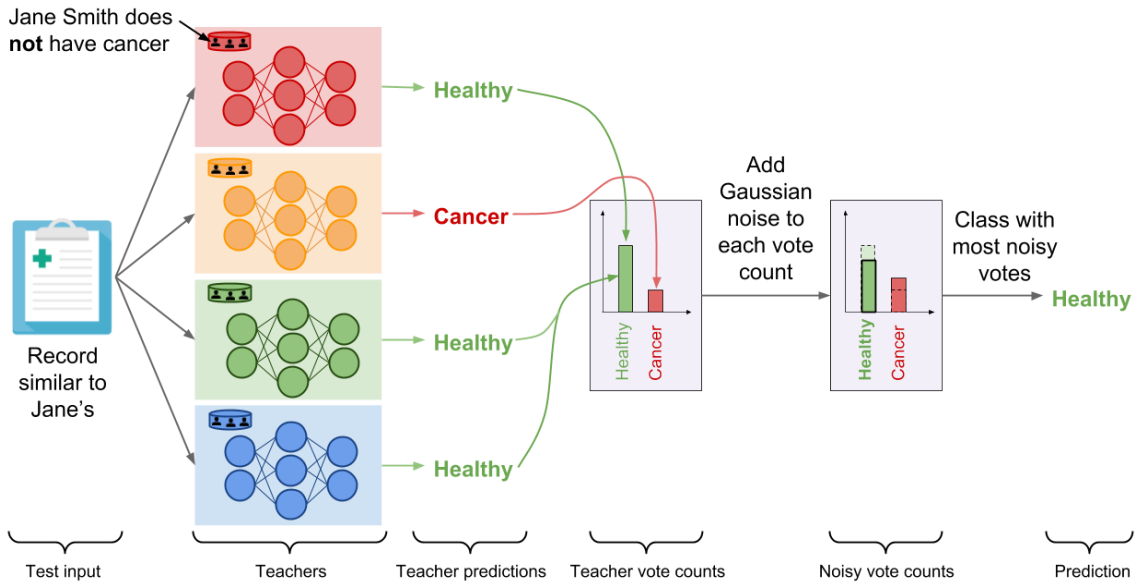
1.4.2.*Differential Privacy* para el machine learning

En este campo, los riesgos a la privacidad emanan de distintos frentes. Por un lado, un modelo puede 'memorizar' partes o la totalidad de ciertos registros. Adicionalmente, un ataque sobre el modelo puede permitir la reconstrucción de los registros originales, exponiendo gran cantidad de información sensible (Goodfellow & Papernot, 2018). Esto último, es especialmente relevante en el contexto de Deep Learning, donde dada la complejidad de los modelos, los mismos son propensos a sobreajustar a los datos y memorizar registros en el límite. Dados estos desafíos, se han propuesto diversas implementaciones de DP en el contexto del aprendizaje automático.

1.4.2.1. *Private Aggregation of Teacher Ensembles*

Una de las implementaciones actualmente más utilizadas es la *Private Aggregation of Teacher Ensembles* - PATE (Goodfellow & Papernot, 2018). Como el nombre sugiere, este método se basa en el entrenamiento de varios modelos de machine learning no privados que se corren sobre subconjuntos diferentes de los datos. Luego, las predicciones de cada modelo son agregadas y finalmente sobre ellas se aplica un mecanismo de DP. Intuitivamente, si diferentes modelos entrenados sobre conjuntos de datos que no comparten ninguna observación entre sí, hacen una misma predicción, esto quiere decir que la misma no depende de la inclusión o no de ningún registro en especial en los datos de entrenamiento. O lo que es lo mismo, este modelo no estará filtrando información sobre ningún registro en especial. El siguiente esquema presenta el funcionamiento de alto nivel de este modelo:

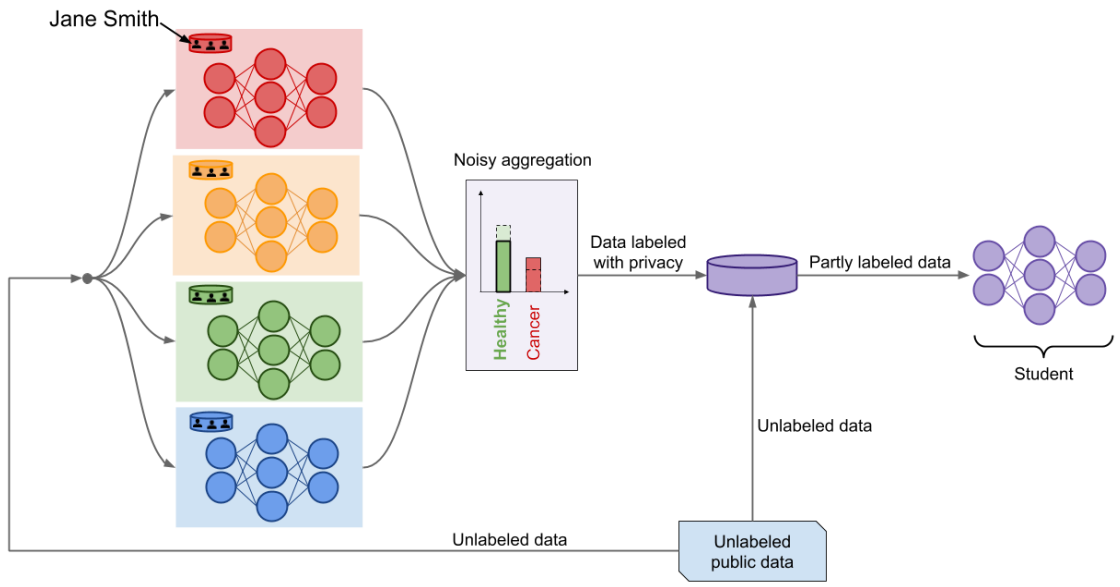
Esquema II. *Private Aggregation of Teacher Ensembles* - PATE - Momento I - Entrenamiento de modelos 'profesores' privados



Fuente: Goodfellow & Papernot (2018)

Si bien a nivel conceptual este modelo como fue presentado funciona y garantiza predicciones privadas, tiene una debilidad: no puede ser publicado ya que los parámetros que aprendió durante el entrenamiento sí pueden revelar información sobre los registros. Por ello, debe complejizarse agregando un paso más. En esta etapa final, se agrega un modelo 'estudiante' que aprende en base a las predicciones de los anteriores. Para ello, toma observaciones aleatoriamente (no utilizadas previamente) sin etiquetar y las hace etiquetar por los modelos anteriores. En base a estas observaciones ya etiquetadas, el modelo estudiante es entrenado. Finalmente, los modelos 'profesores' son eliminados y sólo es utilizado en producción el modelo 'estudiante'. El esquema II presenta esta última etapa del PATE.

Esquema III. *Private Aggregation of Teacher Ensembles* - PATE - Momento II - Entrenamiento del modelo 'alumno'.



Fuente: Goodfellow & Papernot (2018)

1.4.2.2. *Differentially Private Stochastic Gradient Descent* (DP-SGD)

Este modelo implementa garantías de DP durante el entrenamiento de arquitecturas de deep learning. Intuitivamente, lo que hace es aplicar DP en la fase de entrenamiento de los parámetros de la red. Para ello, modifica el algoritmo de descenso estocástico del gradiente utilizado para optimizar los pesos de la red, introduciendo ruido aleatorio en el cómputo del gradiente. El primer paso de este procedimiento es el truncado de los gradientes. Como mencionamos en secciones anteriores, la sensibilidad de una función debe tener una cota superior para que pueda aplicarse DP. Como los gradientes no tienen un límite máximo, los mismos deben truncarse para acotar la sensibilidad. Luego, el algoritmo agrega ruido tomado de una distribución normal sobre el gradiente. Con excepción de estos dos pasos, el resto del proceso de entrenamiento de una red no tiene diferencias respecto al entrenamiento de un modelo no privado (McMahan *et. al.*, 2019).

1.4.3.Differential Privacy para Data Analysis y Data Publishing

1.4.3.1. Data Analysis

Uno de los primeros campos de aplicación en los que comenzó a utilizarse DP es en el ámbito de Data Analysis. Aunque los mecanismos básicos pueden implementarse por usuarios experimentados, ellos revisten cierta dificultad para el uso de parte de usuarios no experimentados. Para superar estos inconvenientes, han sido desarrolladas varias implementaciones de mecanismos de DP para consultas en bases de datos SQL. Entre ellas, destacan los aportes realizados en los frameworks *Sub-Linear Queries* (SuLQ) y *Privacy Integrated Queries Platform* (PINQ). Estos enfoques, lo que permiten es una 'traducción' entre el lenguaje de consultas SQL y los mecanismos de DP (Li *et. al.*, 2017). Si bien estos aportes son básicamente teóricos, han servido de fundamento para desarrollos posteriores como los implementados en la plataforma desarrollada por Microsoft y Harvard, OpenDP, que provee de una API que se conecta a diversos motores de SQL (o compatibles con SQL-92) como `SQL Server`, `PostgreSQL`, `Spark`, `SQLite`, `Pandas`, `Dataverse` y `Presto`. A través de la API, esta plataforma intercepta los requerimientos a la base de datos y devuelve las salidas aplicando los correspondientes mecanismos de DP (Bird, Allen, & Walker, 2020).

1.4.3.2. Data Publishing - Datasets sintéticos

En ocasiones el curador no sólo debe realizar un análisis de los datos, sino que también debe compartir con terceros tal información. De hecho, este es el caso más frecuente y uno de los que más interés recibe en el campo de DP. Existen varias estrategias para compartir la información, y cada una tiene sus ventajas en términos de formato y calidad de la información y privacidad.

Un dataset sintético con garantías de DP es un conjunto de datos de similar estructura al original, pero que contiene registros que han sido generados sintéticamente (Li *et. al.*, 2017). Básicamente, estos datasets mantienen la dimensionalidad de los originales y pueden tener o no (según el criterio del curador), la misma cantidad de registros. El dataset sintético es el formato de publicación que más se asemeja al original. En esencia, un dataset sintético es similar al original, sólo que se le ha aplicado algún mecanismo de DP que ha alterado sus registros.

Justamente porque un dataset sintético mantiene la estructura del original, es la estrategia preferida de publicación de la información, al menos desde la perspectiva de la usabilidad. Como señalan Abowd, Garfinkel & Powazek (2018), el público se halla muy habituado a la publicación de datasets y valora su usabilidad y la flexibilidad que brindan. De todos modos, la generación de datasets sintéticos enfrenta grandes desafíos, principalmente vinculados con el volumen de registros que requieren los algoritmos que se utilizan para generarlos.

A pesar de estas dificultades, este ha sido un campo que recibió bastante atención últimamente. Dada la utilidad (y potencial) de este formato de publicación, desde 2018 el *National Institute of Technology and Standards* ha impulsado diversos concursos abiertos a la comunidad académica para implementar soluciones en el campo de la generación de datasets sintéticos. La primera competencia presentada en 2018 fue el '2018 *The Unlinkable Data Challeng*e' donde se premiaron artículos científicos con propuestas para el desarrollo de datasets sintéticos[19]. Hacia fin de ese mismo año, se agregó a esta competencia el '*Differential Privacy Synthetic Data Challenge*'[20]. En este último concurso, se presentaron diversas soluciones para la generación de datasets sintéticos ya no de modo teórico, sino que en la práctica.

Aunque no son las únicas soluciones propuestas, un enfoque bastante utilizado en torno a la generación de datasets sintéticos es la utilización de redes neuronales. Diversas alternativas se han propuesto en este sentido, muchas de ellas basadas en la implementación de redes recurrentes y de redes generativas (Li *et. al.*, 2017; McKay Bowen & Snoke, 2020). Estos enfoques no sólo están ganando popularidad por la calidad de los resultados alcanzados, sino por el creciente ecosistema que se está gestando en torno a estas soluciones. Una de las bases de estos desarrollos es el referido algoritmo *Differentially private stochastic gradient descent* (DP-SGD) introducido en 2016 por Abadi *et. al,* que permitió incorporar garantías de DP durante el entrenamiento de una red neuronal. Luego, Google decidió incorporar este y otras implementaciones de DP en su librería `Google - Differential Privacy`, que a su vez fueron la base para la librería `Tensorflow Privacy`. La introducción en 2019 de esta librería, que integra los desarrollos previos en torno a DP a uno de los frameworks más utilizados de deep learning, sin duda ha contribuido a popularizar y acercar al público este campo de investigación. Justamente en base a `Tensorflow Privacy` se lanzó en 2020 una librería open source para la generación de datasets

[19]Para mayores detalles, recomendamos revisar el sitio web de la competencia: https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-unlinkable-data-challenge

[20] La información de la primera edición puede consultarse en https://www.challenge.gov/challenge/differential-privacy-synthetic-data-challenge

sintéticos llamada `gretel-synthetics`, desarrollada por la empresa de seguridad y privacidad Gretel.AI.

1.4.3.3. Data Publishing - Tablas de Contingencia

Aunque en general el ideal de Data Publishing sea la publicación de datasets sintéticos, desde la perspectiva de DP esto resulta todavía una tarea desafiante[21]. A pesar de los avances antes mencionados, la construcción de datasets sintéticos es una tarea sumamente compleja y requiere una cantidad de registros muy importante, la que a su vez se incrementa en función de la dimensionalidad del dataset. Por estos motivos, y a pesar del interés que suscita, es una técnica que muchas veces no puede implementarse. Pero incluso si la publicación de microdatos fuese una opción técnicamente viable, hay muchas situaciones en las que se requiere publicar la información de forma agregada. En estas ocasiones, lo que suele realizarse es publicar tablas de contingencia. En general, esta es la estrategia de preferencia de publicación de grandes encuestas, censos, etc.

La tabla de contingencia presenta al público un conjunto de distribuciones conjuntas. El límite superior de distribuciones conjuntas que puede construirse para un dataset de d dimensiones es 2^d. Salvo que estemos ante un dataset de reducida dimensionalidad, esta cantidad de distribuciones conjuntas no tiene sentido de ser publicada. Desde la perspectiva de DP, tal cantidad de potenciales distribuciones conjuntas es un problema, ya que implica una sensibilidad de las consultas muy elevada. A su vez, como el mecanismo de DP agrega ruido a cada punto de la salida, esto significa que va a agregar ruido el orden de 2^d. Ante esta situación, es evidente que el primer desafío para publicar tablas de contingencia, tanto en el contexto de DP, como fuera del mismo, es la delimitación del número de distribuciones conjuntas que se van a publicar (Li *et. al.*, 2017).

Desde la perspectiva técnica, la publicación de tablas de contingencia no implica mayores dificultades, ya que se puede realizar utilizando los mecanismos básicos de DP. No obstante, el desafío para alcanzar salidas de calidad es contar con un volumen de registros importante, ya que en general la sensibilidad de las consultas tiende a ser elevada, y si se cuenta con pocas observaciones, el ruido introducido distorsiona demasiado la información, más allá de un punto razonable.

[21] Al respecto, recomendamos ver Li *et. al*. (2017) y McKay Bowen & Snoke (2020).

1.4.3.4. Data Publishing - Histogramas - Tablas de distribución de frecuencia

En ciertas ocasiones la publicación de tablas de contingencia no es necesaria, o no es posible dada la limitada cantidad de registros en base a los que se trabaja. En este punto, una posibilidad es trabajar en base a la publicación de tablas de distribución de frecuencias o histogramas (que son una representación gráfica de un tipo específico de tabla de distribución de frecuencias para variables numéricas continuas).

Este problema ha sido ampliamente abordado en el contexto de DP, y resulta uno de los más sencillos de encarar. A diferencia del problema anterior de la tabla de contingencia, donde la sensibilidad de la consulta estaba en el orden de 2^d, en el caso de la publicación de tablas de distribución de frecuencias, la sensibilidad será del orden de d. Claro que esta ganancia en términos de reducción de la sensibilidad se logra a costa de publicar menos información. El usuario de las tablas de distribución de frecuencia, ya no puede saber cómo se relacionan las variables entre sí. En cambio, sólo podrá conocer cómo se distribuye cada una de las variables de forma individual (Li *et. al.*, 2017).

Claro que a la hora de elegir una estrategia o formato para publicar la información, no nos encontramos limitados a una única opción. Gracias a la propiedad de la composición presentada anteriormente, pueden aplicarse diversas metodologías para publicar ciertos aspectos de los datos, y luego puede computarse el presupuesto total utilizado. Por ejemplo, se podría publicar un conjunto de tablas de contingencia para exponer la información de ciertas variables de especial interés y luego se puede publicar otra información de menor valor para el usuario en formato de tablas de distribución de frecuencias. Esta estrategia podría completarse con la publicación de algunas estadísticas individuales, como algún conteo o promedios de algunas variables específicas.

1.5. El ecosistema tecnológico en torno a *Differential Privacy*

En base a los desarrollos teóricos sobre DP, durante los últimos años han ido surgiendo una serie de implementaciones de algoritmos y flujos que simplifican el uso de estas tecnologías por parte de usuarios no expertos. En general, se trata de librerías open source que implementan los mecanismos básicos de DP, o extienden las funcionalidades para algún tipo de aplicación específica, como el entrenamiento de modelos de machine learning o consultas en bases de datos SQL. Actualmente, las principales librerías disponibles son las siguientes.

- `diffpriv`:

  Es una librería de R que sólo implementa los mecanismos básicos de DP.

- `google differential-privacy`:

  Es una librería desarrollada por Google disponible en C++, Go y Java. De todos modos, a través de una API en Apache Beam, la librería puede ser utilizada por clientes utilizando otros lenguajes, entre ellos Python. Esta librería tiene implementados los mecanismos básicos de DP y permite realizar consultas como conteo, suma, media, varianza y cuantiles. Asimismo, incluye un módulo para computar el presupuesto y extensiones para realizar consultas sobre bases de datos SQL.

- `PyDP`:

  Es una librería que funciona como un *wrapper* para Python de la librería desarrollada por Google. Permite realizar menos operaciones que su versión original, y al día de hoy, sólo tiene implementado el mecanismo laplaciano.

- `diffprivlib`:

  Es una librería para Python desarrollada por IBM que implementa los mecanismos básicos de DP. Incorpora herramientas para realizar consultas como histogramas, conteos, sumas, medias, cuantiles y varianza. También ofrece un módulo para componer el presupuesto de un conjunto de consultas. Asimismo, brinda funcionalidades para el entrenamiento privado de modelos de machine learning.

- OpenDP - SmartNoise:

  Librería para Python desarrollada a través de un esfuerzo conjunto de Microsoft y Harvard. Implementa los mecanismos básicos de DP y permite realizar cálculos como conteo, suma, media, cuantiles, varianza y covarianza. Entre sus utilitarios, incluye el módulo de cómputo del presupuesto compuesto desarrollado por IBM e implementado en su librería. También ofrece soporte para integración con los motores SQL PostgreSQL y SQL Server y de sintaxis similares basados en Spark, Presto y Pandas. Al igual que diffprivlib, esta librería implementa diversos modelos de machine learning privado, entre los que se incluye también una extensión de PyTorch para entrenar redes neuronales. Asimismo, esta librería incluye un módulo para la generación de datasets sintéticos. A diferencia de otras implementaciones, Microsoft desarrolló una plataforma on - line que se integra con su servicio de cómputo en la nube, Azure, lo que facilita la implementación de mecanismos de DP en flujos de trabajo en la nube.

- Tensorflow Privacy:

  Esta librería también desarrollada por Google implementa mecanismos de DP para el entrenamiento de redes neuronales utilizando Tensorflow.

- Opacus:

  Es una librería de alta velocidad que implementa mecanismos de DP para entrenar redes neuronales utilizando PyTorch

- gretel-synthetics:

  Es una librería desarrollada por la empresa Gretel.AI e implementada en Python que construye sobre Tensorflow y permite la construcción de datasets sintéticos con garantías de DP.

Parte II.

La implementación de la solución

En el marco de los diversos programas de justicia abierta que se hallan en marcha en la Argentina, nosotros estamos colaborando con el Juzgado Penal, Contravencional y de Faltas N° 10 de la Ciudad de Buenos Aires a cargo del Dr. Pablo Casas. Actualmente, las modalidades de publicación de la información judicial utilizadas por el juzgado son dos. Por un lado, se publica on-line el texto de las resoluciones anonimizadas mediante técnicas tradicionales. Sobre este mecanismo de difusión de la información judicial no vemos grandes desafíos a la privacidad (aunque siempre hay aspectos para mejorar)[22]. En cambio, nosotros estaremos trabajando sobre el segundo mecanismo, que efectivamente supone grandes riesgos a la privacidad de los individuos. Además de abrir los textos de las resoluciones, el juzgado publica de forma on-line y con acceso irrestricto un dataset conteniendo datos sobre las causas, que incluye información administrativa y burocrática, de los planteos y los delitos, y de las partes involucradas. La apertura de este conjunto de datos supone varios problemas, que hemos analizado en profundidad en la entrada 1.2.2. del texto. Basta recordar ahora que la publicación on-line de esta información en formato tabular puede permitir la re-identificación de los individuos muy fácilmente, dando por tierra con la expectativa de privacidad de parte de los involucrados.

El objetivo de nuestro trabajo es realizar una propuesta superadora para la apertura de la información contenida en los datasets a los cuales nos referimos más arriba. Para ello, utilizaremos técnicas de DP, de forma tal que no se puedan identificar (más allá de cierto umbral de probabilidad), a los individuos referidos en los registros. Como resultado de la presente investigación, propondremos un flujo de trabajo que permita desarrollar un MVP (*Minimum Viable Product*) que satisfaga las necesidades de apertura de información del juzgado. Dicha propuesta incluirá el detalle de los mecanismos implementados, los parámetros de configuración de los mismos, las transformaciones aplicadas a los datos y un conjunto de recomendaciones de cara a futuras iteraciones.

Esta sección se estructura del siguiente modo. En la primera entrada discutiremos aspectos generales sobre la solución propuesta: evaluaremos las restricciones generales que imponen los datos, presentaremos los supuestos básicos sobre los que trabajaremos y problematizamos las distintas alternativas para aplicar un mecanismo de DP. También analizaremos diversos formatos posibles para la publicación de la información: histogramas, tablas de contingencia y datasets sintéticos. En la segunda entrada haremos un análisis exploratorio de los datos, para entender qué tipo de información contiene el dataset,

[22] Para ver una discusión más profunda al respecto del concepto de privacidad recomendamos revisar la sección 1.1.2.. Asimismo, para una discusión sobre privacidad en el marco de la apertura de datos judiciales, recomendamos referirse a la sección 1.2.2.

propondremos algunos criterios de abordaje para el tratamiento de registros correlacionados y para una primera pre selección de atributos. Asimismo, se procederá a la partición del dataset en sub conjuntos de datos no correlacionados para mejorar el funcionamiento de los mecanismos de DP. En la tercera entrada presentaremos la ingeniería de atributos realizada sobre cada uno de estos subconjuntos de datos. En la cuarta entrada, evaluaremos la *performance* de los mecanismos laplaciano y gaussiano para determinar cuál se aplica mejor a cada subconjunto de datos resultante del procedimiento anterior. También problematizaremos diversos aspectos relativos a la implementación de los mecanismos, como el tratamiento de posibles inconsistencias en las salidas En la quinta entrada analizaremos diversos factores que impactan en la *performance* del mecanismo de DP. Probaremos diferentes estrategias de asignación del presupuesto entre atributos; haremos un análisis de sensibilidad para encontrar un presupuesto óptimo que permita equilibrar privacidad y calidad de la información y estudiaremos la relación entre el tamaño del dataset y la calidad de salida de información. En la sexta entrada recapitularemos el alcance de la propuesta realizada, los límites y los desafíos de cara a futuras implementaciones.

2.1. El mecanismo propuesto

2.1.1. Modalidades de presentación de la información

La primera pregunta que debemos realizarnos es qué tipo de presentación de la información buscamos - o podemos - realizar de acuerdo a las limitaciones técnicas y del conjunto de datos que disponemos. Nosotros consideraremos la utilización de tres estrategias de difusión de la información: la publicación de datasets sintéticos, la publicación de tablas de contingencia, y la publicación de histogramas y/o tablas de distribución de frecuencias. Cada una de estas estrategias brinda diferentes balances entre privacidad y calidad de la información, no obstante, como regla general de acercamiento al problema, podemos decir que cuanta más información deseemos publicar, más compleja será la implementación de un mecanismo capaz de garantizar un balance óptimo entre privacidad y calidad.

2.1.1.1. Datasets sintéticos

Comencemos considerando una estrategia 'ideal' de distribución de los datos: la publicación de datasets sintéticos. Un dataset sintético construido aplicando técnicas de DP es un conjunto de datos tabular que mantiene el formato del conjunto de datos original, pero que incorpora las garantías formales de DP. Como sugieren Abowd, Garfinkel & Powazek (2018), el público se halla muy familiarizado con este tipo de formato y lo prefiere. Adicionalmente, la publicación de datasets conteniendo registros individuales en formato tabular, abre la totalidad de la información disponible al público, con lo que el curador no tiene que presuponer criterios de agregación ni de uso posterior de los datos. Por otro lado, la publicación de microdatos le otorga al usuario final la posibilidad de hacer un aprovechamiento muy flexible de la información, por ejemplo en todo tipo de usos que requieran la utilización de datos ordenados en panel o en formato tabular, como aplicaciones de econometría o aprendizaje automático. En este sentido, la publicación de datasets sintéticos es la estrategia que maximiza la calidad de la información y su usabilidad por parte del público. No obstante, como hemos mencionado en la entrada 1.4.3.2, los desafíos para la publicación de un dataset sintético que resguarde la privacidad de los registros son bastante importantes.

Existen varios enfoques para la publicación de datasets sintéticos. Unos de los más novedosos y prometedores se basan en la implementación de redes neuronales generativas (GAN's) y redes recurrentes (RNN) para la creación de registros artificiales

(McKay Bowen & Snoke, 2020). En el desarrollo de este trabajo, hemos considerado la creación de datasets sintéticos utilizando redes recurrentes, pero los resultados no han sido alentadores. Hemos trabajado con una librería open source desarrollada por Gretel.AI llamada `gretel-synthetics`, que implementa redes recurrentes utilizando el framework de `Tensorflow` y se basa en los desarrollos de Google para implementar los mecanismos de DP[23]. No obstante, los resultados no fueron satisfactorios ya que la calidad de los registros generados fue muy baja. A pesar del intento, estos resultados eran esperables debido a que el tamaño del dataset con el que trabajamos es limitado y tiene gran dimensionalidad. Adicionalmente, el tokenizador que utiliza el algoritmo considera caracteres independientes, lo que puede ser útil para predecir letras y finalmente palabras, pero que no parece funcionar bien para predecir categorías[24], al menos en el contexto de un mecanismo de DP. Al introducir ruido, el mecanismo de DP da lugar a predicciones de caracteres sin sentido. En cambio, fuera del contexto de DP la misma librería es capaz de generar registros sintéticos de muy buena calidad.

2.1.1.2. Tablas de contingencia

En segundo lugar en términos de usabilidad-calidad de la información podemos ubicar la publicación de tablas de contingencia. En general, y salvo ciertas excepciones, como es la Encuesta Permanente de Hogares[25] (EPH) en la Argentina, los institutos de estadística utilizan casi exclusivamente esta estrategia de publicación. Aunque no suelen aplicar técnicas de DP, hace décadas que en la disciplina de estadística son conscientes del riesgo a la privacidad que supone publicar la totalidad de la información distribuyendo microdatos de censos y encuestas (Dalenius, 1977; Abowd, Garfinkel & Powazek, 2018)[26]. Por ello, este tipo de publicación fue progresivamente abandonado en favor de publicaciones de datos agregados, incluyendo tablas de contingencia, histogramas y tablas de distribución de frecuencias. La riqueza de la tabla de contingencia es que permite analizar distribuciones conjuntas y captar correlaciones entre variables. Aunque cierto tipo de usuarios prefieran la utilización de microdatos, las tablas de contingencia son un recurso muy valioso en términos de la calidad de información que distribuye. No obstante, la producción de éstas requiere un

[23] Para la utilización de esta librería nos hemos basado en las publicaciones de https://gretel.ai/blog/how-to-create-differentially-private-synthetic-data y https://medium.com/gretel-ai/using-generative-differentially-private-models-to-build-privacy-enhancing-synthetic-datasets-c0633856284

[24] Nuestro conjunto de datos está constituido por variables categóricas principalmente.

[25] La EPH abre todos los microdatos que recopila en sus encuestas y los publica en formato tabular. Aunque esto aporta gran riqueza para el análisis científico, también supone un gran riesgo a la privacidad de los encuestados.

[26] El Censo de los Estados Unidos de 2020 está siendo procesado aplicando técnicas de DP. De hecho, cuando sus resultados estén publicados va a ser el primer censo de gran escala en ser procesado de este modo. Al respecto sugerimos leer *National Academies of Sciences, Engineering, and Medicine* (2021).

esfuerzo extra de parte del curador. A diferencia del caso anterior, donde el curador simplemente publicaba el conjunto de datos y se desentendía del procesamiento posterior, en este caso debe decidir los criterios de agregación de la información y más específicamente qué tablas construir. Esto supone tanto conocimiento del dominio, como antelación a las necesidades de los usuarios. A su vez, la publicación de tablas de contingencia impone cierta rigidez para el uso posterior de los datos, ya que algún tipo de análisis que el usuario final quiera realizar, y que no fue considerado previamente por el curador, puede verse restringido por cómo la información fue expuesta en las tablas.

En el campo de DP la publicación de tablas de contingencia es una práctica bastante estudiada. No obstante, no por ello está exenta de obstáculos. El principal problema radica en la cantidad de distribuciones conjuntas que pueden publicarse. En teoría, el límite superior de distribuciones conjuntas que existen es 2^d. No es difícil suponer que esta es una cantidad inabordable de distribuciones conjuntas, incluso al margen de las limitaciones impuestas por los mecanismos de DP. Por ello, uno de los principales desafíos es la determinación de cuáles son las distribuciones conjuntas que tiene sentido publicar. Recordemos que para cualquier mecanismo de DP, la adición del ruido depende principalmente de dos parámetros, del presupuesto y de la sensibilidad de la consulta. En el caso de una tabla de contingencia, lo que tenemos son tantas consultas correlacionadas, como distribuciones conjuntas publiquemos. Si medimos la sensibilidad global de este grupo de consultas, ya sea que utilicemos la norma L1 o L2, sus valores serán muy elevados (salvo que publiquemos muy poca información). En el caso de utilizar la norma L1, la sensibilidad global de la consulta estará en el orden de 2^d, mientras que si utilizamos la norma L2 estará en el orden de $\sqrt{2^d}$. Entonces, el problema que se genera para aplicar DP a la publicación de tablas de contingencia, es que el ruido a adicionar es muy elevado, con lo que la calidad de la información se deteriora mucho (Li *et. al.*, 2017).

De todos modos, existen algunas alternativas y lineamientos para poder utilizar este mecanismo de publicación de información. Un punto de partida es la recopilación de la mayor cantidad de registros posibles. Como el ruido a agregar a cada punto de la salida es de una magnitud acotada[27], el impacto que tiene sobre conteos relativamente bajos es mucho mayor que el que tiene sobre conteos más elevados. Por lo tanto, siempre que se trabaje con mecanismos de DP el primer punto a considerar será conseguir tantos registros

[27] Como mencionamos más arriba, el ruido a agregar a cada punto de salida de la consulta depende únicamente del presupuesto asignado y de la sensibilidad de la consulta. Con estos parámetros, se construye una distribución con media 0 y se toma muestra aleatoria de la misma. Por ello, no podemos decir que la magnitud del ruido agregar es fija, sino que lo que es fija es la distribución de la que se toma dicho valor.

como sea posible, como para disminuir la distorsión relativa introducida por el mecanismo utilizado. En segundo lugar, deben seleccionarse cuidadosamente las distribuciones conjuntas que vayan a ser publicadas, de forma tal de disminuir la sensibilidad global de la consulta a valores razonables. En general, lo que se intenta realizar para mejorar la calidad de la información publicada es incrementar el tamaño del dataset y disminuir la sensibilidad de la consulta.

En vista a estos condicionamientos que impone la publicación de tablas de contingencia, consideramos que no resultan adecuados para nuestra aplicación. Básicamente, ello se debe a lo reducido del dataset con el que contamos y a su elevada dimensionalidad. Es decir, tenemos muchas distribuciones conjuntas potenciales para publicar, y muy pocas observaciones para cada una de ellas. Eso supone dos cosas. Primero una gran sensibilidad de las consultas. Segundo, al haber conteos bajos, el ruido a agregar desnaturalizaría completamente la información, al punto de inutilizarla. Pero, incluso aunque se pudiesen salvar algunos de estos aspectos a través de la publicación de pocas tablas de contingencia cuidadosamente seleccionadas, todavía debemos evaluar que ello haga sentido. Aunque es algo que no estudiamos para nuestro caso por las limitaciones antes señaladas, tampoco podemos asumir directamente que la publicación de tablas de contingencia aportará una calidad informativa muy superior a lo que son métodos más sencillos de publicación. De todos modos, de cara a futuras iteraciones de este trabajo nos gustaría explorar con más detalle la viabilidad de esta estrategia para publicar la información.

2.1.1.3. Histogramas - Tablas de distribución de frecuencias

La publicación de histogramas o tablas de distribución de frecuencias procesados con DP es la estrategia de distribución de la información más sencilla, y supera muchas de las limitaciones que estuvimos señalando hasta este momento. En primer lugar, la sensibilidad global de las consultas es muy inferior al caso de la publicación de tablas de contingencia. En el caso anterior hablábamos de un rango de sensibilidad entre el orden de $\sqrt{2^d}$ usando la norma L2 y de 2^d usando la norma L1. Para la publicación de histogramas, la sensibilidad estará entre el orden de $\sqrt{d}$ y d, de acuerdo a la norma que se utilice para medirla (Li ***et. al.***, 2017; Fathima, 2020). La ventaja sobre las otras estrategias de publicación en este sentido es innegable. La magnitud del ruido a agregar a las consultas es bastante menor que con otros métodos, por lo que la calidad de la información resultante será mayor. Esto hace que sea la estrategia ideal (o la única viable a veces), en situaciones donde existe una

cantidad limitada de registros. Aunque la disminución en la cantidad de registros impacta negativamente en la calidad de la información al igual que en los casos anteriores, el umbral mínimo para alcanzar salidas de calidad aceptable es mucho más bajo.

A pesar de esta gran ventaja que tiene la publicación de histogramas, este formato deja de lado también una gran cantidad de información. Por ejemplo, no captura correlaciones entre variables. El usuario de un histograma no tiene forma alguna de entender cómo se relacionan entre sí las variables que tiene delante. Otro resguardo a tener en cuenta respecto a los histogramas, es que a pesar de que tiendan a tener una sensibilidad menor que las tablas de contingencia, esto no exime del control de la dimensionalidad de la información a distribuir. El curador sigue siendo responsable de elegir cuidadosamente qué se publicará, ya que si intenta publicar todo el conjunto de datos disponible, en casos de datasets de alta dimensionalidad, el ruido a agregar tornará la información una vez más inutilizable. Por último, debe señalarse que la publicación de histogramas-tablas de distribución de frecuencias, es sensible a la cardinalidad de los atributos (lo mismo vale para las tablas de contingencia). Aunque la cardinalidad no influye en la medida de sensibilidad y no guarda relación con el ruido a adicionar, sí guarda relación con la calidad de la salida de la información. Atributos con una gran cardinalidad y menos registros en cada categoría, tenderán a sufrir más el agregado de ruido que atributos con menos categorías (Li. *et. al.*, 2017).

La correcta selección-reorganización de las categorías para el caso de variables categóricas, como de rangos para el caso de variables numéricas, es tarea prioritaria del curador. De hecho, para la publicación de histogramas propiamente dichos existen una familias de algoritmos como el `NoiseFirst` y el `StructureFirst`, que están especialmente diseñados con esta problemática en mente, que es la determinación de la cantidad y el rango de los bines a publicar (Li. *et. al.*, 2017).

Dadas las alternativas anteriores repasadas, **para nuestra tarea hemos elegido continuar trabajando en base a la publicación de tablas de distribución de frecuencias**. Consideramos que dadas las limitaciones antes enumeradas es la única alternativa viable. De todos modos, como fue mencionado previamente, no descartamos de cara al futuro seguir complejizando la propuesta y considerar otras estrategias, como la publicación de tablas de contingencia.

2.1.2. Completando el *setting* inicial

La información en las tablas de frecuencias será publicada como una distribución de probabilidades normalizada (donde la suma de las frecuencias para cada categoría de una variable suma 1), lo que creemos que simplificará la lectura de parte del público y favorecerá la utilización de mecanismos de DP basados en distribuciones continuas (como el laplaciano y el gaussiano).

Otro punto importante a señalar es que el flujo que estamos proponiendo está pensado para aplicarse en *batches* de datos, todos independientes entre sí. Por la naturaleza de su actividad, el juzgado está siempre incorporando registros al conjunto de datos. Existen varias alternativas para abordar esta dinámica. Hay mecanismos para trabajar con información que arriba en *streams*, como también para considerar la superposición parcial de batches (Li *et. al.*, 2017). No obstante, estas técnicas están por fuera del alcance de este abordaje. Para nuestra propuesta, estaremos considerando un único conjunto de datos sin ningún tipo de relación con otros posibles conjuntos que se generen posteriormente. Naturalmente, la evolución de este desarrollo es la consideración de futuros incrementos de registros en el dataset, no obstante, eso lo dejaremos para otro trabajo.

Por último, recordamos que para evaluar la calidad de las salidas construidas, nos valdremos de las métricas presentadas en la sección 1.3.7. Para ello, utilizaremos como medidas de error el EAM (Error Absoluto Medio) y la divergencia KL. En el anexo II de la tesis puede consultarse el procedimiento seguido para implementar estas medidas de error en el contexto de los mecanismos de DP.

2.1.3. Las herramientas a utilizar

Para el tratamiento de los datos y la implementación de las diversas soluciones de DP, utilizaremos el ecosistema de herramientas disponible en torno a Python. En el análisis y transformación de los datos, haremos uso principalmente de las facilidades provistas por `Pandas`. En cuanto a las soluciones de DP, trabajaremos con la librería de IBM para `Python diffprivlib`, a la que le hemos realizado algunas pequeñas modificaciones para adaptarla a nuestra metodología de trabajo. Esta librería de IBM es una herramienta de código abierto que implementa diversos mecanismos de DP y herramientas, entre las que destacan módulos para la construcción de histogramas, para el cómputo del presupuesto y diversas implementaciones de algoritmos de aprendizaje automático privado. Actualmente

(abril de 2021), son dos las librerías que implementan esta diversidad de herramientas: `OpenDP - SmartNoise`, apadrinada por Microsoft y la universidad de Harvard y la ya mencionada de IBM. Aunque en implementación ambas librerías son bastantes similares, la de IBM fue la primera en lanzarse al mercado, motivo por el cual es la que estaremos utilizando. Esto no implica que de cara a futuras iteraciones, no pueda evaluarse la conveniencia de trabajar con `OpenDP`.

2.2. Explorando el dataset

El set de datos que estaremos abordando abarca todas las resoluciones que hacen a las causas que trató el Juzgado Penal, Contravencional y de Faltas 10 de la Ciudad de Buenos Aires entre agosto de 2016 y noviembre de 2020. Entre ellas, destacan los registros ligados a violencia de género. De un total de 3853 registros, 1069 corresponden a casos de violencia de género. Asimismo, el juzgado le ha dado gran relevancia a la difusión de estos casos, por el interés social, político y científico que reviste. El resto de los casos que se tramitan involucran una variedad importante de delitos, desde faltas y contravenciones hasta ofensas contra el código penal de la nación, pasando por *habeas corpus* y ejecuciones de multas.

Este dataset contiene información relativa a las resoluciones dictadas por el juzgado. **Cada registro corresponde a un planteo diferente. Un planteo es una acusación específica, de una actividad que transgredió (supuestamente) un artículo de una ley o código. Cada resolución puede contener más de un planteo**, **con lo cual, una misma resolución, puede incluir varios registros distintos**. Es decir, en el mismo episodio, un individuo puede estar violando a la vez diferentes artículos de una ley o diferentes leyes. Del mismo modo, en relación a una única acusación puede existir más de un planteo. **Aunque todos ellos son evaluados en el marco de una única resolución, la misma se compone de diferentes planteos, sobre los que el juez se expide posteriormente de forma individual**. Más allá del significado jurídico de estas figuras, lo que nos interesa a nosotros en el marco de la implementación de una estrategia de DP es que vamos a tener registros correlacionados, y eso es un obstáculo que más adelante veremos cómo es abordado.

En relación al contenido de los registros, cada uno presenta datos relativos al planteo. Para entender en profundidad qué tipo de información tienen estos registros, agruparemos los campos en base a algunos criterios que nos permitirán simplificar el abordaje. Por un lado tenemos información relativa al delito-planteo. Por otro lado, tenemos datos relativos a la víctima y al agresor. Para los casos de violencia de género también existe información sobre la naturaleza de la relación entre agresor y víctima y del hecho en sí. Luego, tenemos información sobre la decisión del juez. Otro conjunto de atributos se refiere a aspectos burocráticos-administrativos respecto al planteo y la causa. Por último, hay información relativa a la apelación de la causa, tanto en su definición como a aspectos administrativo-burocráticos. En total, el conjunto de datos tiene 61 atributos (en esta cuenta

no se considera la columna **N**) que engloban todos estos puntos. Una descripción detallada de los mismos puede hallarse en el anexo I de la tesis.

2.2.1. La estructura del dataset - Preselección de atributos

Este conjunto de datos contiene 3853 registros y 61 atributos[28] , que incluyen 1069 registros vinculados a violencia de género y 2785 correspondientes a otras causas. Incluso antes de comenzar a estudiar en profundidad el dataset ya detectamos que tenemos muchas dimensiones para un dataset de tamaño reducido. Recordemos de la sección anterior lo que planteamos sobre la publicación de histogramas: a mayor cantidad de dimensiones d, el ruido a incorporar a fin de garantizar DP es mayor, y la calidad de la salida se tiende a deteriorar. Por ello, debemos priorizar en base a los requerimientos del caso, qué información será publicada. A continuación haremos un repaso por los atributos del dataset, revisando la información que contienen y priorizando cuáles serán conservados para proseguir el trabajo.

En la tabla II tenemos todos los atributos del dataset y el tipo de información que contienen (no se incluye la columna **N,** que es la clave primaria y será utilizada como índice del dataframe durante la manipulación de los datos).

Tabla II. Atributos agrupados según tipo de información.

Atributo	**Tipo de información**
MATERIA	Delito
ART_INFRINGIDO	
CODIGO_O_LEY	
CONDUCTA	
CONDUCTA_DESCRIPCION	
MODALIDAD_DE_LA_VIOLENCIA	
ZONA_DEL_HECHO	
FRECUENCIA_EPISODIOS	Delito - Violencia de Género
RELACION_Y_TIPO_ENTRE_ACUSADO/A_Y_DENUNCIANTE	
HIJOS_HIJAS_EN_COMUN	
MEDIDAS_DE_PROTECCION_VIGENTES_AL_MOMENTO_DEL_HECHO	

[28] El dataset tiene 61 atributos, pero nosotros sólo consideraremos 60, ya que uno es la clave primaria.

LUGAR_DEL_HECHO	
VIOLENCIA_DE_GENERO	
V_FISICA	
V_PSIC	
V_ECON	
V_SEX	
V_SOC	
V_AMB	
V_SIMB	
V_POLIT	
FRASES_AGRESION	
GENERO_DENUNCIANTE	Víctima
NACIONALIDAD_DENUNCIANTE	
EDAD_DENUNCIANTE_AL_MOMENTO_DEL_HECHO	
NIVEL_DE_INSTRUCCION_DENUNCIANTE	
GÉNERO ACUSADO/A	Agresor
NACIONALIDAD_ACUSADO/A	
EDAD_ACUSADO/A_AL_MOMENTO_DEL_HECHO	
NIVEL_DE_INSTRUCCION_ACUSADO/A	
TIPO_DE_RESOLUCION	Resolución
OBJETO_DE_LA_RESOLUCION	
DETALLE	
DECISION	
ORAL_ESCRITA	
N	Burocrático-Administrativo
NRO_REGISTRO	
FECHA_RESOLUCION	
FIRMA	
HORA_DE_INICIO	
HORA_DE_CIERRE	
LINK	
DURACION	
SI_NO_RECURRENTE	Apelación
DECISION_CAMARA_DE_APELACIONES	

N_REGISTRO_Y_TOMO_CAMARA	
LINK_CAMARA	
SI_NO_RECURRENTE_CAMARA	
DECISION_DE_ADMISIBILIDAD_CAMARA	
N_REGISTRO_Y_TOMO_CAMARA_1	
LINK_CAMARA_1	
QUEJA_Y_RECURRENTE	
DECISION_DE_ADMISIBILIDAD_TSJ	
N_REGISTRO_Y_TOMO_TSJ	
LINK_TSJ	
DECISION_DE_FONDO_TSJ	
N_REGISTRO_Y_TOMO_TSJ_1	
LINK_TSJ_1	
RECURSO_EXTRAORDINARIO_Y_RECURRENTE	
DECISION_CSJN	
N_REGISTRO_Y_TOMO_CSJN	
LINK_CSJN	

Si bien los criterios con los que se agruparon los atributos en función del tipo de información que contienen son en última instancia arbitrarios - y susceptibles de ser reordenados -, nos van a servir para ir jerarquizando aquellos que revisten información de interés. Antes de profundizar en el detalle de estos grandes grupos, revisemos algunos campos que merecen especial atención, ya sea porque funcionan como claves primarias o foráneas o porque consisten en enlaces a otros documentos:

- **N**: es la clave primaria del dataset.
- **NRO_REGISTRO**: es la clave de cada causa. Vincula los diferentes planteos entre sí. Lo utilizaremos más adelante para hacer un análisis de registros correlacionados.
- **LINK**: es el link al texto de la resolución del juzgado.
- **N_DE_REGISTRO_Y_TOMO_CAMARA:** número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones.
- **LINK_CAMARA:** es el link al texto de las resoluciones de la Cámara de Apelaciones del Fuero.
- **N_DE_REGISTRO_Y_TOMO_CAMARA_1:** número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones

- **LINK_CAMARA_1:** link al texto de la resolución de la cámara de apelaciones del fuero.
- **N_DE_REGISTRO_Y_TOMO_TSJ:** número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones del fuero
- **LINK_TSJ**: es el link al texto de la resolución de la cámara de apelaciones del fuero.
- **N_DE_REGISTRO_Y_TOMO_TSJ_1:** número de registro interno del Juzgado para las resoluciones del Tribunal Superior de Justicia.
- **LINK_TSJ_1**: es el link al texto de la resolución del tribunal superior en caso de apelación.
- **N_REGISTRO_Y_TOMO_CSJN:** número de registro de la resolución de la Corte Suprema, en caso de existir.
- **LINK_CSJN:** es el link al texto de la resolución del Tribunal Superior de Justicia en caso de existir.

Estos campos no poseen información que tenga sentido ser abierta al público (no tiene lógica hacer una tabla de distribución de frecuencias de claves primarias, como tampoco pueden agregarse datos como los links a las resoluciones). Con excepción del campo **NRO_REGISTRO**, que nos permitirá hacer un estudio de correlaciones entre registros, el resto de los campos pueden ser completamente dejados de lado para lo que sigue del trabajo.

Examinemos a continuación los grupos de atributos que establecimos anteriormente. Es bastante claro que los atributos relativos a aspectos burocráticos-administrativos de las causas, van a resultar de poco interés para el público. No queremos decir que carezcan completamente de valor, para algún tipo de estudio puede resultar de valor saber cuál fue la duración en tiempo de una audiencia, o las fechas de las resoluciones, para, por ejemplo, analizar la estacionalidad de las mismas. No obstante, dado que tenemos que priorizar qué información distribuir, consideramos que este conjunto de atributos puede ser dejado de lado sin mayores inconvenientes.

Ahora debemos jerarquizar y elegir del resto de información que conservamos, cuál mantendremos - aunque sea de forma provisoria - para su publicación. Responder esta pregunta nos lleva a reexaminar y problematizar los objetivos de apertura de la información de parte del juzgado. Habíamos dicho que seguían dos objetivos prioritarios. Por un lado, dar publicidad a los actos de gobierno. Esto ya nos deja saber que cierta información relativa a las decisiones habrá de conservarse. Por otro lado, sabemos que el juzgado está interesado en abrir hacia el público en general, y especialmente hacia científicos e

investigadores, todo tipo de información que pueda contribuir al incremento del conocimiento colectivo. Esto implica la publicación de información relativa a las características de los delitos, de las víctimas y de los agresores, con especial énfasis en los casos de violencia de género. En lo que sigue de nuestro análisis, nos guiaremos en base a estas prioridades.

Siguiendo estos lineamientos, un subconjunto de atributos que ya podemos dejar de lado son los que refieren a las apelaciones de las causas, ya que no refieren a actos de gobierno del propio juzgado, ni informan sobre características de los hechos. Si bien esta información no es irrelevante, en el marco de la jerarquización que estamos realizando consideramos esta información como secundaria.

Considerando el descarte de variables realizado hasta el momento, ya hemos reducido bastante la dimensionalidad del conjunto de datos. De todos modos, necesitamos reducir aún más la dimensionalidad del dataset a publicar. Para ello, re examinaremos los atributos que conservamos hasta el momento, a ver si existen otros que podamos descartar. El estudio de los datos que viene a continuación se basa en el trabajo realizado en el apartado 1.1 de los notebooks, donde referimos al lector para consultar los detalles del análisis sobre cada variable.

Del conjunto de atributos relativos a las características del delito, vamos a descartar la columna **ZONA_DEL_HECHO**. Si bien este dato es interesante, el nivel de agregación que presenta en el dataset hace que no sea verdaderamente portador de información de calidad. También descartaremos la columna **MATERIA**, ya que la misma se halla correlacionada con otros atributos relativos a la caracterización del delito, como la ley y los artículos infringidos (ya que la materia de la competencia depende principalmente de estas variables - aunque no exclusivamente). Luego, del conjunto de atributos relativos a violencia de género, destacaremos los siguientes: **V_ECON**, **V_PSIC**, **V_SEX**, **V_SOC**, **V_AMB**, **V_SIMB** y **V_POLIT**. Como vemos en tabla III donde analizamos este conjunto de variables asociadas a violencia de género, estos atributos tienen baja varianza, con lo que el aporte de información que pueden hacer es bajo, al menos en relación con los otros atributos que estamos conservando.

Tabla III. Varianza de atributos seleccionados.

Atributo	Varianza	Atributo	Varianza
V_FISICA	0.,248	V_SEX	0,070
V_ECON	0,168	V_PSIC	0,005
V_SOC	0,126	V_SIMB	0,003
V_AMB	0,096	V_POLIT	0

Asimismo, la columna **FRASES_AGRESION** será descartada, ya que este tipo de información tampoco puede ser publicada en el formato que utilizaremos. También eliminaremos las columnas **GENERO_ACUSADO/A** y **GENERO_DENUNCIANTE** de los grupos de datos de agresor y denunciante respectivamente. En ambos casos, estos atributos tienen poca varianza y alta cardinalidad. Asimismo, el género del denunciante se carga únicamente para los casos de violencia de género.

De los conjuntos restantes de atributos, eliminaremos la columna **ORAL_ESCRITA**, correspondiente al grupo de atributos relativos a la resolución, ya que consideramos que su valor relativo es bajo. Terminada esta primera selección de atributos, hemos eliminado 38 atributos y conservamos 22 para seguir trabajando.

2.2.2. Análisis de registros correlacionados

Como se explicó en la entrada 1.3.5., en el contexto de DP se entiende que son registros correlacionados todos aquellos que puedan estar vinculados a un único individuo y consecuentemente, pueden ser afectados por la inclusión o exclusión del mismo en el conjunto de datos. Por ejemplo, en una base de datos de ventas, todas aquellas ventas que refieran a un mismo cliente van a estar correlacionadas.

En casos donde existen registros correlacionados existen dos alternativas. La primera es aplicar PD utilizando técnicas de *Group Privacy*. Básicamente, esto implica agregar mayor ruido en la utilización de los mecanismos, de forma tal de enmascarar no ya la inclusión de un individuo en los registros, sino de todo el grupo. Esto tiene dos obstáculos. Primero, que debe establecerse una cota máxima a la cantidad de registros correlacionados que puede tener el dataset, ya que la misma entrará en juego a la hora de determinar la sensibilidad de las consultas. De no existir tal cota, debe establecerse arbitrariamente y descartarse todo registro correlacionado que exceda tal umbral. La segunda dificultad radica en que el nivel de ruido a agregar para garantizar DP con registros correlacionados, crece en función de la cota máxima establecida. Más allá de que se utilice la norma L1 o la norma L2 en la

determinación de la sensibilidad del mecanismo, el ruido adicionar crece muy rápidamente. En el mejor de los escenarios, utilizando la sensibilidad L2, la existencia de un número tan bajo como 4 registros correlacionados estará duplicando la cantidad de ruido a adicionar por el mecanismo. La alternativa a trabajar con *Group Privacy* es descartar los registros correlacionados y conservar una única observación por grupo. De este modo, aunque se pierde información por el descarte de registros, es probable que la calidad de la salida sea mejor que utilizando *Group Privacy*, ya que se adiciona menos ruido.

Como expusimos en secciones anteriores, en el caso de nuestro dataset sabemos que tenemos registros correlacionados. Cada registro corresponde a un planteo y una resolución puede tener varios planteos[29]. Así, tenemos múltiples registros que refieren a un mismo individuo, con lo que todos estos registros deben considerarse correlacionados.

Tabla IV. Distribución de frecuencias de planteos por resolución

Cantidad de planteos por resolución	Conteo de resoluciones	Cantidad de planteos por resolución	Conteo de resoluciones
1	2504	5	8
2	414	6	3
3	113	7	0
4	27	8	2

Como se puede verificar en la tabla IV, casi todos las resoluciones contienen un único planteo, pero existen varias que contienen más de uno. De hecho, hay varias causas con 4, 5, 6 y hasta 8 planteos. Tenemos, para 3853 planteos, 3071 resoluciones. Hecha esta verificación, existen dos posibilidades para continuar el abordaje: utilizar *Group Privacy* o descartar los registros correlacionados y utilizar DP de forma tradicional. Cualquier alternativa tiene sus costos en términos de pérdida de calidad de la información y complejidad de implementación.

Para evaluar la conveniencia de los diferentes enfoques, haremos un análisis de sensibilidad del error en función de distintas variantes de implementación. Se comparará la calidad de la información de salida utilizando DP de forma tradicional descartando todos los registros correlacionados y también aplicando DP bajo el concepto de *Group Privacy*, con diferentes cotas superiores para la cantidad máxima de registros correlacionados (y descartando los registros que superan esta cota máxima). De este modo, la medida de error a la que llegaremos en este experimento dará cuenta tanto del ruido introducido por la DP, como el

[29]Para un análisis detallado de la estructura del dataset, recomendamos referirse a la sección 2.2 del trabajo.

efecto del descarte aleatorio de observaciones correlacionadas. El desarrollo de este experimento puede consultarse en la entrada 1.2 de los notebooks, donde referimos para revisar el detalle de la implementación realizada.

Para el ejercicio actual, aplicaremos DP utilizando el mecanismo laplaciano (más adelante probaremos con otros). Lo que haremos, será construir tablas de distribución de frecuencias aplicando DP para todas las variables del dataset (que no han sido descartadas en la etapa previa). Para limitar el número de los registros correlacionados, en los casos que verifiquemos que superen la cota máxima establecida, seleccionaremos de forma aleatoria los registros que serán considerados y descartaremos los otros.

Antes de comenzar con el experimento, nos quedan por determinar dos parámetros. El primero es el presupuesto de privacidad. Como simplemente estamos haciendo un análisis de sensibilidad del efecto de la inclusión de registros correlacionados sobre el error, lo dejaremos fijo en un valor arbitrario. A tales fines, hemos establecido un presupuesto de privacidad de 1 ε para cada tabla de distribución de frecuencias (construiremos una por cada variable del dataset). Resta aún evaluar el tema de la sensibilidad de la función. Recordemos de la entrada 1.3.3. que la sensibilidad L1 de un histograma o tabla de distribución de frecuencias es es 1. De aquí se desprende que para la construcción de k tablas, la sensibilidad L1 también es de k. Como tenemos 21 dimensiones sobre las que aplicaremos DP (y construiremos una tabla de distribución de frecuencias para cada una de ellas), la sensibilidad total será también de 21 (recordemos que no aplicamos DP sobre la columna **NRO_REGISTRO**). Para garantizar la robustez de los resultados, realizaremos 250 ejecuciones y reportaremos el error promedio. Los que siguen son los resultados del experimento realizado en base a estos criterios:

Gráfico IV. Análisis de sensibilidad del error en función de la cantidad de registros correlacionados. Resultados de 250 ejecuciones.

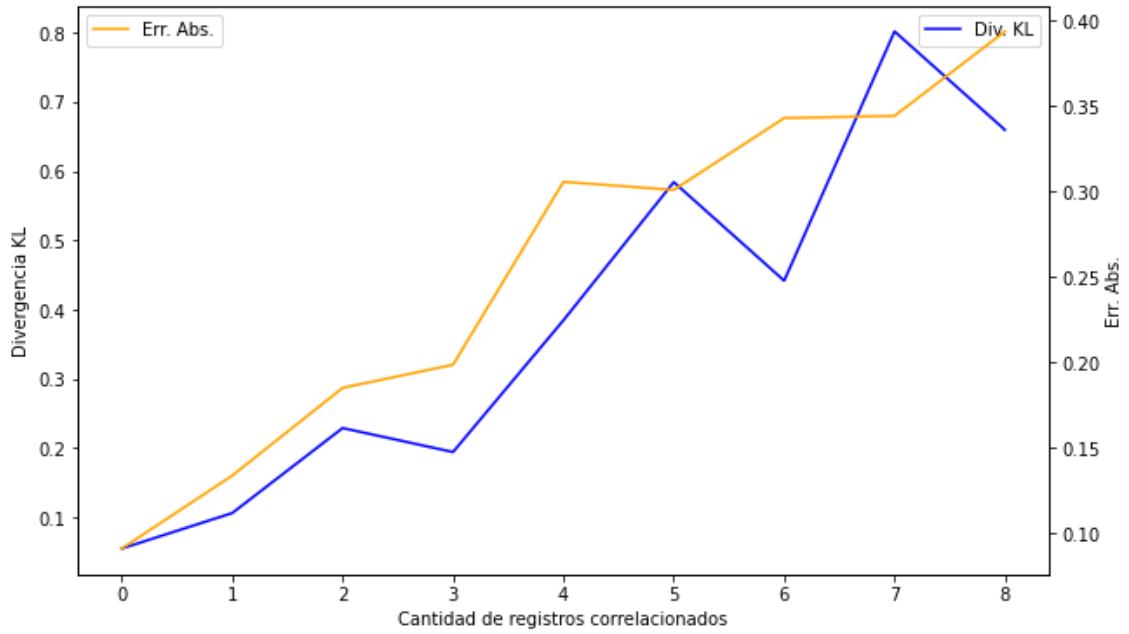


Tabla V. Análisis de sensibilidad del error en función de la cantidad de registros correlacionados. Resultados de 250 ejecuciones.

Cantidad de registros correlacionados	Div. KL	Error Absoluto
0	0.054677	0.091246
1	0.105939	0.133872
2	0.228776	0.185102
3	0.19407	0.198614
4	0.384406	0.305479
5	0.584172	0.300782
6	0.441698	0.342741
7	0.801874	0.34402
8	0.659632	0.393362

La salida del experimento señala que en este caso, las métricas de error son mejores trabajando sin registros correlacionados, ello a pesar de que estamos descartando 782 registros, alrededor de un 20% del total. Esto se debe a que el deterioro sobre la información es más importante debido al incremento del ruido necesario para trabajar con *Group Privacy*, que debido al descarte de registros. El error inducido por el descarte de registros

es mínimo, en cambio, el error inducido por el mecanismo de DP representa casi la totalidad del mismo. Como comentamos anteriormente, el ruido a agregar en un contexto de *Group Privacy* crece en función de la cantidad de registros correlacionados. Por ello, el deterioro de la información progresa muy rápidamente, mucho más que a causa del descarte de registros.

En base a estos resultados, decidimos seguir trabajando sin registros correlacionados, ya que es la alternativa que mejor calidad de información ofrece en nuestro caso. Para ello, en las resoluciones que tengan más de un planteo (cada registro del dataset corresponde a un planteo), elegiremos de forma aleatoria uno solo y descartaremos el resto[30]. Luego de realizado este procedimiento, el tamaño del dataset quedará reducido a 3071 registros, uno por resolución.

2.2.3. Partición del dataset

Como último paso de esta etapa de definiciones de abordaje, particionaremos el dataset en dos: uno incluirá todos los registros vinculados a violencia de género y el otro el resto de las causas. De ahora en más, nos referiremos al conjunto de datos que conserva los registros vinculados a causas de violencia de género como set de datos I y al conjunto de datos que contiene al resto de los registros, como set de datos II.

Estos dos subconjuntos tienen características muy diferentes por cómo está estructurado el dataset originalmente. De hecho, los registros sobre violencia de género tienen muchos atributos particulares, que en el caso de las otras causas son completados con valores nulos o "No corresponde". Por otro lado, una columna que deja de tener sentido particionando el set de datos, es **VIOLENCIA_DE_GENERO**, ya que su varianza se verá reducida a 0 para cada sub conjunto de datos, con lo que puede ser descartada directamente, ayudando a reducir la dimensionalidad de los datasets resultantes.

Adicionalmente, al particionar el dataset, evitaremos el uso de tablas de contingencia, lo que podría haber sido interesante o necesario para entender la distribución de cada atributo en relación a los casos de violencia de género. De este modo, reducimos la complejidad de la solución, ya que cada tipo de causas es tratado por separado. Por otro lado, podremos tratar los atributos de la forma que sea más conveniente de acuerdo al tipo de causa.

[30] En el apartado 2.2.1 hemos expuesto en detalle la estructura del dataset.

Este abordaje se valdrá de la ventaja derivada de la composición paralela de las consultas, como fue presentado en la entrada 1.3.6.1. El presupuesto total utilizado para asegurar DP no va a ser la suma de los presupuestos aplicados en ambos datasets, como sería si los enfocamos como consultas secuenciales, sino el mayor presupuesto aplicado a cualquiera de ellos.

2.3. Análisis e ingeniería de atributos

En el apartado anterior hemos planteado estrategias generales de abordaje de los datos. Hicimos un primer descarte de atributos en función de la información que priorizamos para publicar. Los atributos que mantuvimos son 21, aunque como usaremos el campo **NRO_REGISTRO** como índice del dataframe, nos quedarán 20 campos para seguir trabajando. También definimos una estrategia para resolver el obstáculo de los registros correlacionados, conservando únicamente un registro por resolución. Por último, particionamos el dataset en dos sub conjuntos, uno conteniendo las causas de violencia de género (dataset I), con 795 registros y otro conteniendo el resto de las causas (dataset II) con 2276 registros.

En lo que sigue del apartado presentaremos las principales decisiones adoptadas en cuanto a ingeniería de atributos, presentando el análisis de bajo nivel sólo cuando la complejidad o el desarrollo de la explicación lo ameriten. Por lo demás, solo haremos referencias a las transformaciones realizadas. Para consultar los detalles de cada transformación, sugerimos al lector remitirse a los notebook's de la tesis.

El objetivo del análisis e ingeniería de atributos que realizaremos en esta etapa será terminar de definir si existe algún feature que deba descartarse[31] y la cardinalidad óptima para cada variable. Recordemos que la dimensionalidad del dataset se vincula con la sensibilidad de la consulta y la cardinalidad de la variable tiene impacto en la cantidad de ruido total a adicionar, y consecuentemente en la calidad de la salida.

2.3.1. Cardinalidad de la variable y calidad de la información

Antes de continuar con la presentación de la ingeniería de atributos, detengámonos un poco en el problema de la cardinalidad de la variable y su relación con el deterioro de la calidad de la información. Imaginemos que estamos agregando ruido a dos variables, una con dos categorías, y la otra con cinco. Para simplificar el ejercicio, diremos que la primera variable tiene 500 observaciones en cada categoría y la segunda variable tiene 200 observaciones en cada categoría. Luego, a cada variable le aplicamos DP utilizando un mecanismo gaussiano, que (a los fines de este ejemplo) va a adicionar ruido tomado de una $\mathcal{N}(0,1)$[32]. Aplicando las propiedades de una normal, sabemos que la distribución resultante para cada

[31] Por ejemplo, por tener muchos valores faltantes, alta entropía o baja varianza.
[32] Esta distribución fue propuesta para simplificar el ejemplo, nótese que no hemos tenido en cuenta la sensibilidad ni el delta para determinar la distribución de la que tomamos el ruido.

categoría será la suma de un escalar y de la distribución anterior. Para el caso de la categoría con 500 observaciones, la nueva distribución será una $\mathcal{N}(500, 1)$, mientras que para la categoría con 200 observaciones la nueva distribución será una $\mathcal{N}(200, 1)$. Rápidamente podemos concluir que la adición de ruido tomada de variables aleatorias independientes e idénticamente distribuidas (IID) distorsiona mucho más conteos bajos, que conteos altos. Para el primer caso, el coeficiente de variación de la distribución resultante de la adición del ruido es de 1/500, mientras que para la segunda es de 1/200. A medida que la cantidad de observaciones disminuye, el impacto de la adición del ruido tiende a acrecentarse (Li *et. al.*, 2017). Por ello, debemos analizar los atributos, a fin de reestructurar las categorías del mejor modo posible, tratando de reducir la cardinalidad, pero también evitando perder información.

2.3.2. Ingeniería de atributos para el set de datos I

2.3.2.1. Atributos con valores missings y de poca varianza

Comenzaremos estudiando si existen atributos portadores de poca información que podamos descartar. Lo primero que buscaremos es descartar atributos con muchos valores faltantes. Consideramos como valores faltantes aquellos que tienen campos en blanco, o las leyendas ‘s\d’ y ‘no_corresponde’. En la tabla VI vemos la proporción de observaciones que cumplen esta definición de valores faltantes. Los detalles de este procesamiento pueden seguirse en el apartado 2.1.1 de los notebooks.

Tabla VI. Valores faltantes por atributo.

Atributo	**Porcentaje de valores faltantes**
ART_INFRINGIDO	0%
CODIGO_O_LEY	0%
CONDUCTA	0%
CONDUCTA _DESCRIPCION	38%
V_FISICA	0%
MODALIDAD_DE_LA_VIOLENCIA	13%
NACIONALIDAD_ACUSADO/A	4%
EDAD_ACUSADO/A AL MOMENTO DEL HECHO	4%

NIVEL_INSTRUCCION_ACUSADO/A	21%
NACIONALIDAD_DENUNCIANTE	4%
EDAD_DENUNCIANTE_AL_MOMENTO_DEL_HECHO	21%
FRECUENCIA_EPISODIOS	3%
RELACION_Y_TIPO_ENTRE_ACUSADO/A_Y_DENUNCIANTE	2%
HIJOS_HIJAS_EN_COMUN	2%
MEDIDAD_DE_PROTECCION_VIGENTES_AL_MOMENTO_DEL_HECHO	2%
LUGAR_DEL_HECHO	2%
TIPO_DE_RESOLUCION	0%
OBJETO_DE_LA_RESOLUCION	0%
DETALLE	10%
DECISION	0%

Aquí vemos que la única columna que parece tener faltantes en exceso es **CONDUCTA_DESCRIPCION**, con casi el 40% de valores faltantes. Las columnas **NIVEL_INSTRUCCION_ACUSADO/A** y **EDAD_DENUNCIANTE_AL_MOMENTO_DEL_HECHO** también tienen valores altos, superiores al 20% de faltantes, pero resultan de interés para análisis posteriores. Por ello, por ahora sólo descartaremos la columna **CONDUCTA_DESCRIPCION**.

2.3.2.2.Combinación de atributos

Una vez que eliminamos los atributos con valores faltantes, vamos a evaluar la posibilidad de combinar algunos de los atributos restantes en una única variable y luego reordenar sus categorías. Para ello, revisaremos la distribución conjunta de algunas variables que sabemos (o sospechamos) que pueden estar correlacionadas. Así, podremos determinar si tiene sentido combinar los dos atributos en uno nuevo. Comenzamos estudiando la distribución conjunta de las variables **CODIGO_O_LEY** y **ART_INFRINGIDO**.

Tabla VII. Distribución conjunta de las variables CODIGO_O_LEY y ART_INFRINGIDO

Coordenada (CODIGO_O_LEY / ART_INFRINGIDO)	Frecuencia	Porcentaje acumulado de obs.
codigo_penal_de_la_nacion / 149bis	326	41%
codigo_contravencional / 52	138	58%
ley_13944 / 1	81	68%
codigo_penal_de_la_nacion / 92	47	74%
codigo_contravencional / 53bis_inc5	32	79%
codigo_penal_de_la_nacion / 239	26	82%
codigo_penal_de_la_nacion / 89	23	85%
codigo_penal_de_la_nacion / 189bis	15	86%
ley_24270 / 1	14	88%
codigo_contravencional / 67bis	13	89%

La tabla VII muestra para cada coordenada de la distribución conjunta la frecuencia absoluta y la proporción acumulada de observaciones. Puede observarse que sólo 5 coordenadas representan más del 75% de las observaciones. A fin de reducir la dimensionalidad del dataset, podemos combinar estos dos atributos en uno nuevo que se llamará **LEY_Y_ARTICULO,** y reducirlo a 6 niveles.

Un procedimiento similar se ha seguido con la distribución conjunta de las variables **MODALIDAD_DE_LA_VIOLENCIA** y **LUGAR_DEL_HECHO**. En la tabla VII podemos observar su distribución conjunta.

Tabla VIII. Distribución conjunta de las variables MODALIDAD_DE_LA_VIOLENCIA y LUGAR_DEL_HECHO

Coordenada (MODALIDAD_DE_LA_VIOLENCIA / LUGAR_DEL_HECHO)	Frecuencia	Porcentaje acumulado de obs.
domestica / en_domicilio_particular	387	55%
domestica / en_domicilio_particular_y_mediante_medios_tecnologicos	75	66%
domestica / via_publica	67	76%
domestica / mediante_medios_tecnológicos	62	85%

domestica / puerta_de_domicilio_particular	31	90%
domestica / en_domicilio_laboral	12	91%
domestica / via_publica_y_en_domicilio_particular_y_mediante_medios_tecnológicos	12	93%
domestica / s/d	11	94%
domestica / en_domicilio_particular_y_en_puerta_de_domicilio_particular	7	95%
domestica / en_auto_particular	6	96%

Al igual que en el caso anterior, vemos que unas pocas coordenadas acumulan el grueso de las observaciones. Estos dos atributos pueden ser combinados en uno nuevo, que llamaremos **MODALIDAD_Y_LUGAR DEL HECHO**. Este nuevo atributo tendrá 5 niveles.

2.3.2.3. Análisis y reestructuración de la cardinalidad de las variables

Luego de terminar el análisis de distribuciones conjuntas en los casos considerados relevantes, pasaremos al estudio y transformación de los niveles de cada variable. Como hemos mencionado más arriba, nuestro objetivo es reordenar las categorías de las variables de tal modo que resulte la menor cantidad posible, pero sin perder información relevante. Para ello, nos quedaremos con las categorías más frecuentes, que agrupen al menos el 75% de las observaciones (intentando mantener siempre la cantidad de categorías en no más de 5, de ser posible) y reagruparemos el resto de las variables en una nueva categoría. En los casos donde las variables tengan alta cardinalidad y las observaciones se hallen muy dispersas entre las categorías, desistiremos de trabajar con ellas y las descartaremos. En dichos casos, no es posible reordenar las categorías sin corromper la naturaleza de los datos, por lo que resulta más conveniente descartar la variable en su conjunto.

A continuación presentaremos la información sumarizada para todas las variables, indicando si fue descartada o no, la cantidad de niveles originales y la cantidad de niveles luego de la transformación (si no se realizó ninguna transformación la cantidad de niveles será igual en ambos casos). Para consultar los detalles del análisis y criterios seguidos sobre cada variable, referimos al lector al apartado 2.1.2 de los notebooks.

Tabla IX. Sumario de las transformaciones aplicadas a cada variable.

Variable	N° Cat. Originales	N° Cat. Reord.	Descartada
CONDUCTA	29	5	No
CONDUCTA_DESCRIPCION	18	0	Sí
V_FISICA	3	3	No
NACIONALIDAD_ACUSADO/A	16	5	No
EDAD_ACUSADO/A AL MOMENTO DEL HECHO *	6	6	No
NIVEL_INSTRUCCION_ACUSADO/A	13	5	No
NACIONALIDAD_DENUNCIANTE	12	5	No
EDAD_DENUNCIANTE_AL_MOMENTO_DEL_HECHO *	6	6	No
FRECUENCIA_EPISODIOS	6	5	No
RELACION_Y_TIPO_ENTRE_ACUSADO/A_Y_DENUNCIANTE	12	4	No
HIJOS_HIJAS_EN_COMUN	3	3	No
MEDIDAS_DE_PROTECCION_VIGENTES_AL_MOMENTO_DEL_HECHO	3	3	No
TIPO_DE_RESOLUCION	2	2	No
OBJETO_DE_LA_RESOLUCION	40	7	No
DETALLE	114	0	Sí
DECISION	2	2	No
CODIGO_O_LEY	4	0	Si (Combinada)
ART_INFRINGIDO	32	0	Si (Combinada)
MODALIDAD_DE_LA_VIOLENCIA	6	0	Si (Combinada)
LUGAR_DEL_HECHO	17	0	Si (Combinada)
LEY_Y_ARTICULO	0	6	Nueva Variable
MODALIDAD_Y_LUGAR DEL HECHO	0	5	Nueva Variable

* Variables continuas discretizadas

En la tabla IX constatamos que hemos aplicado profundas transformaciones en los atributos. De los 20 atributos iniciales, terminamos con 16. Pero las principales modificaciones fueron

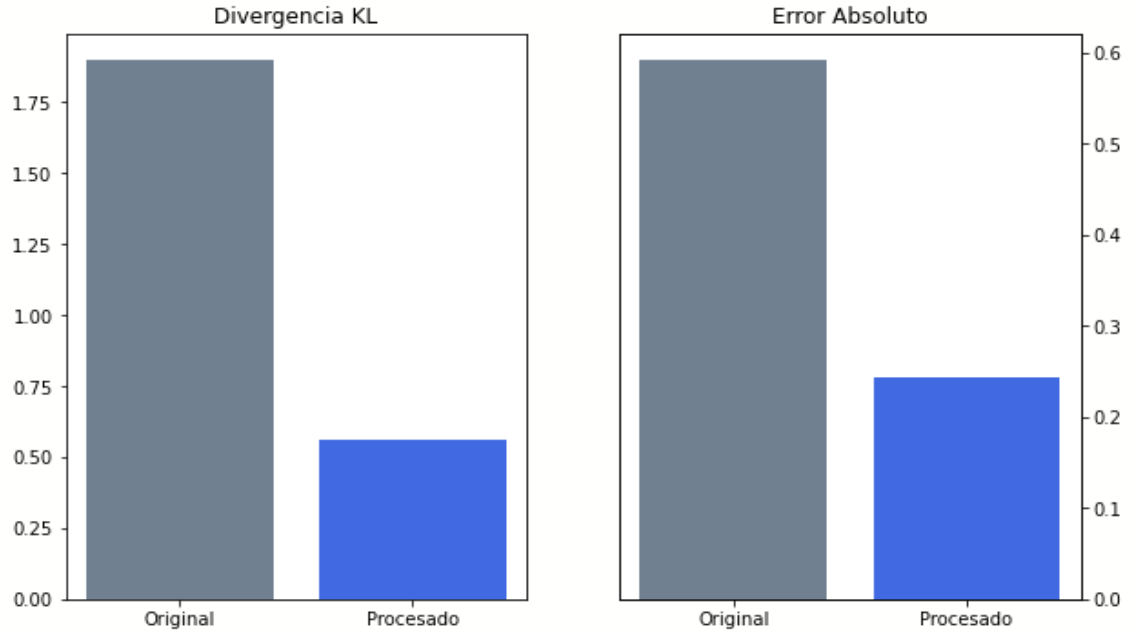
aplicadas a nivel de la cardinalidad de las variables. Inicialmente, teníamos 344 niveles considerando todos los atributos (las variables continuas las consideramos como si hubiesen estado ya discretizadas), al finalizar el procesamiento, terminamos con 71 niveles. Es decir, reducimos en promedio la cardinalidad de los atributos en un 80%. Incluso dejando de lado en este cálculo la variable **DETALLE**, que tenía una cardinalidad de 114, la reducción promedio sigue siendo elevada, superior al 65%.

2.3.2.4. Evaluación de las transformaciones introducidas

Hasta este punto nos hemos guiado por la teoría para aplicar transformaciones que potencialmente podrían mejorar la calidad de salida de la información. Veamos a continuación si realmente estas modificaciones implicaron mejora alguna en la calidad de salida de los mecanismos de DP. Para analizar esto, aplicaremos DP sobre el dataset I, con todos los atributos originales y sus respectivos niveles. Luego, aplicaremos DP sobre el conjunto de datos ya transformado y compararemos las métricas de error entre ambos conjuntos de datos. Para la aplicación del mecanismo de DP, utilizaremos un presupuesto de 5 ε, que será distribuido entre todos los atributos. En esta prueba, utilizaremos el mecanismo laplaciano, que es el mecanismo por default para la construcción de tablas de distribución de frecuencias. Respecto a la sensibilidad de la función, recordemos los desarrollos de apartados anteriores, donde señalamos que para estas consultas, debemos considerar una sensibilidad equivalente a la cantidad de atributos sobre los que vayamos a construir tablas de distribución de frecuencias[33]. Repetiremos la aplicación del mecanismo 100 veces para llegar a conclusiones más robustas y reportaremos el promedio de los resultados. En la entrada 2.1.3 de los notebooks de la tesis puede seguirse en detalle este desarrollo.

[33] Una discusión sobre el problema de la sensibilidad puede encontrarse en la entrada 1.3.3. de la tesis.

Gráfico V. Métricas de error promedio del mecanismo de DP antes y después de aplicar ingeniería de atributos sobre el dataset I. Resultados de 100 ejecuciones.



Esta salida indica que las transformaciones introducidas en este conjunto de datos redundaron en mejoras en la calidad de la información. Ambas métricas de error presentan una disminución superior al 50%. Para el caso de la divergencia KL, esta pasó de 1,90 sobre el dataset original, a 0,56 luego de realizada la ingeniería de atributos, mientras que el error absoluto pasó de 0,59 a 0,24. Esta mejora obedece a dos efectos. Por un lado, al disminuir la cantidad de atributos disminuyó la sensibilidad global de la función. No obstante, la mayor ganancia proviene de haber reordenado la cardinalidad de las variables.

2.3.3. Ingeniería de atributos para el set de datos II

Hasta aquí estuvimos repasando la ingeniería de atributos realizada sobre el conjunto de datos con resoluciones relativas a casos de violencia de género. Sigamos a continuación con el conjunto de datos que contiene el resto de las causas.

2.3.3.1.Atributos con valores missings y de poca varianza

Al igual que con el conjunto de datos anterior, trataremos de descartar atributos con proporciones elevadas de valores faltantes. Recordemos que consideramos como faltantes los campos en blanco y con valores ‘s/d’ y ‘no_corresponde’. En la tabla X vemos la proporción de observaciones que cumplen esta definición de valores faltantes. Los detalles de este procesamiento pueden seguirse en el apartado 3.1.1 de los notebooks.

Tabla X. Valores faltantes por atributo.

Atributo	Porcentaje de valores faltantes
ART_INFRINGIDO	1%
CODIGO_O_LEY	1%
CONDUCTA	0%
CONDUCTA _DESCRIPCION	77%
NACIONALIDAD_ACUSADO/A	20%
EDAD_ACUSADO/A AL MOMENTO DEL HECHO	24%
NIVEL_INSTRUCCION_ACUSADO/A	47%
LUGAR_DEL_HECHO	3%
TIPO_DE_RESOLUCION	0%
OBJETO_DE_LA_RESOLUCION	0%
DETALLE	7%
DECISION	0%

En base a este análisis, descartaremos los atributos **CONDUCTA _DESCRIPCION,** ya que tiene mayoría de valores faltantes **y NIVEL_INSTRUCCION_ACUSADO/A,** que cuenta con casi un 50% de valores faltantes**.** Los campos **NACIONALIDAD_ACUSADO/A** y **EDAD_ACUSADO/A_AL_MOMENTO DEL HECHO** también tienen proporciones de faltantes elevadas, pero son atributos de interés, por los que no los descartaremos.

2.3.3.2. Combinación de atributos

Del mismo modo que se realizó con set de datos I, para este conjunto de datos, analizaremos la distribución conjunta de las variables **CODIGO_O_LEY** y **ART_INFRINGIDO.** Así, definiremos la posibilidad de combinar estas dos variables en una nueva a fin de reducir la dimensionalidad del dataset. El detalle del procesamiento realizado puede seguirse en la entrada 3.1.1 de los notebooks de la tesis.

Tabla XI. Distribución conjunta de las variables CODIGO_O_LEY y ART_INFRINGIDO

Coordenada (CODIGO_O_LEY / ART_INFRINGIDO)	Conteo	Porcentaje acumulado de obs.
codigo_penal_de_la_nacion / 128	191	8%
codigo_penal_de_la_nacion / 149bis	132	14%
codigo_contravencional / 73	113	19%
codigo_contravencional / 118	111	24%
codigo_contravencional / 111	100	29%
codigo_penal_de_la_nacion / 189bis	92	33%
codigo_contravencional / 74	87	37%
codigo_contravencional / 76	83	40%
codigo_contravencional / 114	81	44%
codigo_contravencional / 97	81	47%
ley_23737 / 14	54	50%
ley_23737 / 5c	54	52%
ley_451 / 6.1.94	52	55%
codigo_penal_de_la_nacion / 183	42	56%
codigo_contravencional / 95	42	58%
codigo_penal_de_la_nacion / 181_inc1	38	60%
ley_451 / 6.1.52	35	61%
codigo_contravencional / 52	33	63%
codigo_contravencional / 86	30	64%
codigo_contravencional / 88	29	66%

A diferencia de lo que pasaba en el conjunto de datos I, donde las observaciones se acumulaban en unas pocas coordenadas, en este caso vemos una gran dispersión. Ello tiene sentido, ya que este conjunto de datos incluye causas muy diversas y no parece haber un predominio claro de algún tipo en especial. Dada esta dispersión, no tiene sentido combinar estos dos atributos en uno nuevo, ya que de hacerlo deberíamos crear muchos niveles, que es justamente una de las cosas que deseamos evitar. Lo que haremos

entonces será continuar trabajando con el atributo **CODIGO_O_LEY** (más adelante veremos si conviene redefinir su cardinalidad), y descartaremos el atributo **ART_INFRINGIDO**, ya que esta variable, fuera de contexto del código a la ley, no tiene sentido en sí misma.

2.3.3.3. Análisis y reestructuración de la cardinalidad de las variables

A fin de reordenar la cardinalidad de los atributos, sobre este dataset repetiremos el procedimiento utilizado en el dataset I. Al igual que en el caso anterior, nuestro objetivo es reducir en lo posible la cantidad de niveles existentes, pero sin perder información relevante. Para ello, conservaremos las categorías más frecuentes de cada variable que incluyan al menos el 75% de las observaciones y reagruparemos las restantes, tratando de reducir los niveles existentes a no más de 5 (llegando a 7 en como máximo). Las variables que por su elevada cardinalidad no puedan ser reorganizadas en sus categorías, serán descartadas.En la tabla XII presentamos la información sumarizada de todas las transformaciones introducidas en el dataset, indicando la cantidad de niveles de cada atributo antes y después de las transformaciones y si fue conservado o descartado. Este procesamiento puede consultarse en detalle en la entrada 3.1.2 de los notebooks de la tesis.

Tabla XII. Sumario de las transformaciones aplicadas a cada variable.

Variable	N° Cat. Originales	N° Cat. Reord.	Descartada
CONDUCTA	128	0	Sí
CONDUCTA_DESCRIPCION	36	0	Sí
NACIONALIDAD_ACUSADO/A	29	3	No
EDAD_ACUSADO/A AL MOMENTO DEL HECHO *	6	6	No
NIVEL_INSTRUCCION_ACUSADO/A	17	0	Sí
TIPO_DE_RESOLUCION	2	2	No
OBJETO_DE_LA_RESOLUCION	57	7	No
DETALLE	199	0	Sí
DECISION	2	2	No
CODIGO_O_LEY	13	4	No
ART_INFRINGIDO	150	0	Sí
LUGAR_DEL_HECHO	22	6	No

* Variables continuas discretizadas

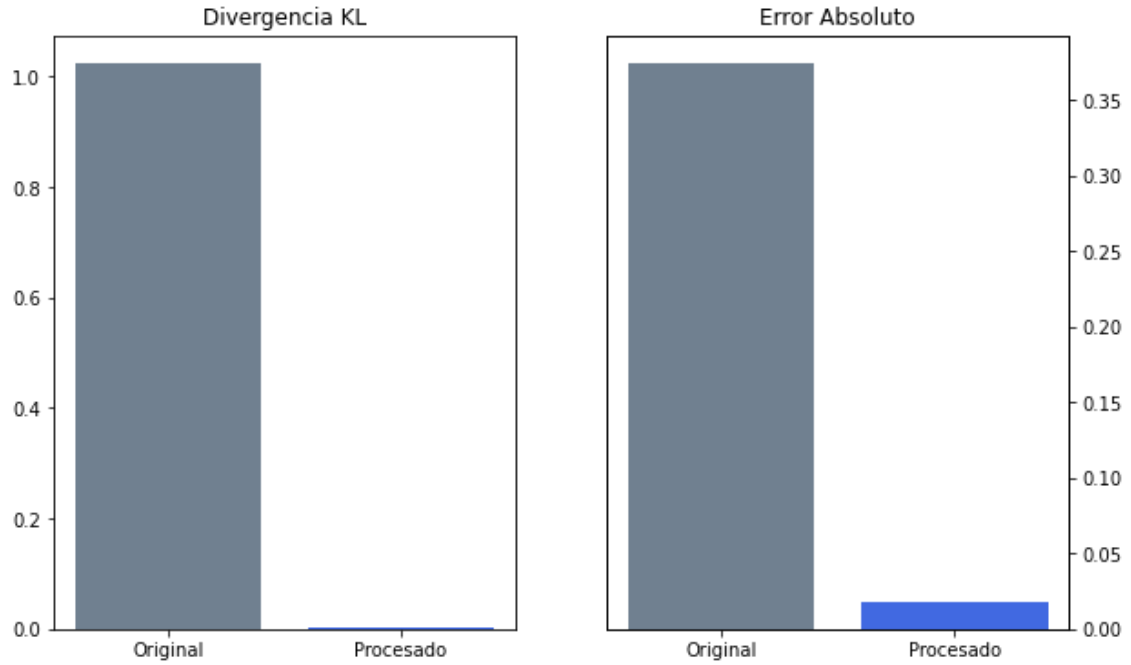
De los 12 atributos originales, hemos descartado 5 y conservamos 7. Este descarte, sumado a la reducción en la cardinalidad de las variables conservadas, permitió reducir la cantidad de categorías iniciales de 661 a las 30 finales.

Para este conjunto de datos hemos aplicado un descarte de atributos más grande que el que realizamos con el conjunto anterior. Ello se debe a que tenemos más columnas con valores faltantes, pero también a que existe una dispersión importante en la información de las causas, lo que lleva a que no podamos reducir a unos pocos niveles la cardinalidad de muchas variables. Esta situación lo que sugiere es que la información tal como está expuesta actualmente en el dataset, es de poca utilidad. El simple hecho de contar con 661 categorías para un dataset de este tamaño, da cuenta de la dificultad para generalizar que posee esta información. Sería recomendable pensar en nuevas formas de organizar estos datos, de cara a un uso más provechoso de la misma.

2.3.3.4. Evaluación de las transformaciones introducidas

Del mismo modo que evaluamos las transformaciones que realizamos sobre el dataset anterior, lo haremos aquí. Compararemos el error de aplicar DP sobre el dataset antes de las transformaciones, con el error de aplicar DP, luego de las mismas. Para la configuración del mecanismo de DP, utilizaremos el mecanismo laplaciano, con un presupuesto de 5 ε que se repartirán entre todos los atributos, y la sensibilidad de la función estará dada por la cantidad de atributos sobre los que apliquemos DP. Al igual que en el caso anterior, repetiremos estos cálculos 100 veces para garantizar resultados robustos y reportaremos los promedios del error. Este procedimiento puede revisarse en la entrada 3.1.3 de los notebooks de la tesis.

Gráfico VI. Métricas de error promedio del mecanismo de DP antes y después de aplicar ingeniería de atributos sobre el dataset II. Resultados de 100 ejecuciones.



Lo que nos permite ver este experimento es que el error introducido por el mecanismo de DP luego de aplicada la ingeniería de atributos en este conjunto de datos, es muy bajo, prácticamente de 0. La divergencia KL pasó de 1,02 a 0, mientras que el error absoluto disminuyó de 0,37 a 0,02. Si bien esto da cuenta de la conveniencia del reordenamiento de la cardinalidad de las variables, consideramos que refleja principalmente el descarte de atributos, que alcanzó a 5 de las 12 columnas originales. Si bien existe un *trade off* entre el descarte de variables "problemáticas" y la calidad global de la salida de la información, en este caso es evidente que la información original era muy deficiente y el descarte, necesario. Ello puede observarse en la magnitud del error absoluto de la salida del mecanismos de DP sobre el dataset original, que es superior al 35%. Publicar un conjunto de datos con un error absoluto del 35% (con una divergencia KL también alta), no tiene sentido desde el uso posterior que se pueda realizar de esta información, ya que la distribución original de la misma ha sido bastante desnaturalizada. En este escenario, es preferible publicar menos información pero de mayor calidad, a intentar publicar toda la información disponible a riesgo de que sea de poca calidad y usabilidad.

2.4. Comparación de mecanismos de *Differential Privacy*

En esta entrada del trabajo, estaremos comparando dos mecanismos de DP - el laplaciano y el gaussiano - a fin de determinar cuál funciona mejor en los conjuntos de datos que tenemos y dadas las consultas que pretendemos realizar. Esta comparación la haremos por separado para cada sub conjunto de datos.

2.4.1. Revisitando el concepto de *Differential Privacy*

En esencia un mecanismo de DP lo que hace es agregar ruido a la salida de una consulta sobre un conjunto de datos. El funcionamiento básico de tal mecanismo M puede representarse como:

$$M \ = \ Consulta(db) \ + \ ruido$$

Asimismo, este mecanismo debe cumplir con la garantía formal de DP. Recordemos la formalización de DP que presentamos en la entrada 1.3.2 de la tesis.

> Un algoritmo aleatorio M con dominio $\mathbb{N}^{|x|}$ es (ε, δ) - *differentially private* para todas las salidas $S \subseteq Rango(M)$ y para todo $x, \ y \in \mathbb{N}^{|x|} \ tal \ que \ ||x - y||_1 \leq 1$:
>
> $$\Pr\ [M(x) \in S] \leq e^{\varepsilon} \Pr\ [M(y) \in S] \ + \ \delta$$
>
> Donde:
>
> S: representa todas las salidas posibles del algoritmo M.
> x : representa todas las entradas en la base de datos original.
> y: representa todas las entradas de la base de datos paralela (con n - 1 entradas).
> ε: representa el presupuesto de privacidad.
> δ: representa la probabilidad de un evento adverso de goteo de datos.

Básicamente lo que nos señala esta inecuación, es que la diferencia (proporcional) en las probabilidades condicionales de obtener determinada salida de una consulta, dado que se corrió sobre x o sobre y debe ser menor a e^{ε} + δ. Recordemos, que esta restricción vale para todos los pares de puntos entre la base de datos original x y todas las bases de datos paralelas y posibles. Esta restricción, otorga un límite teórico a la cantidad de información que un adversario puede ganar respecto a la inclusión de determinada observación en la

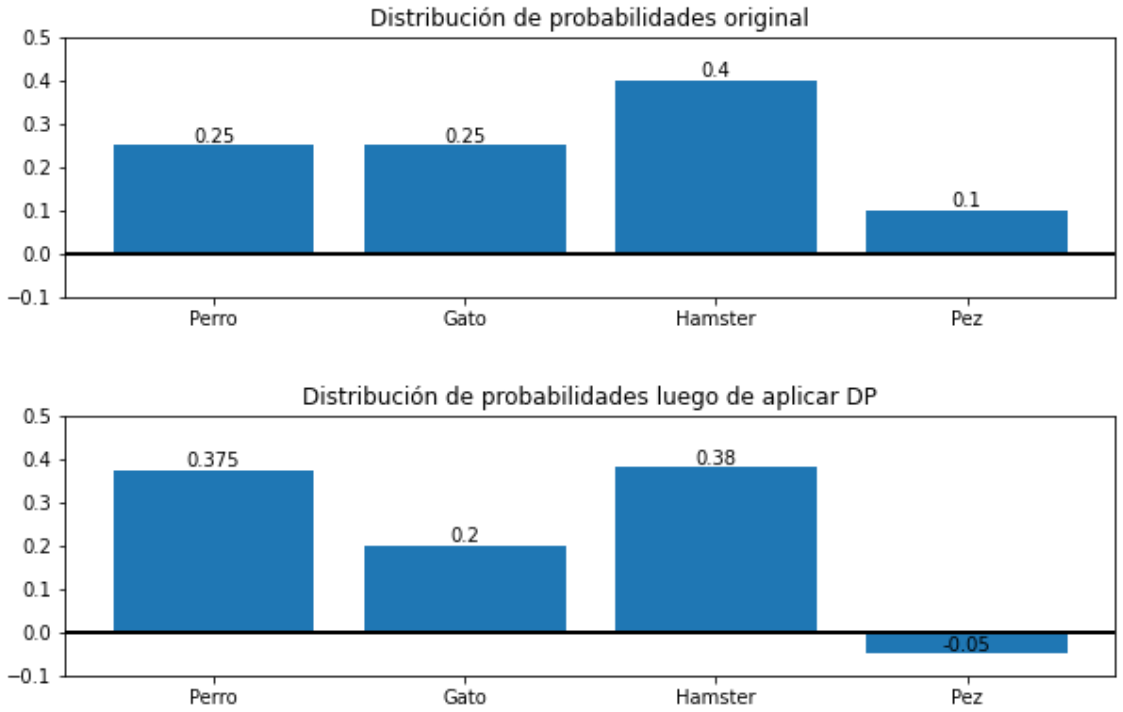
base de datos. Cualquier mecanismo que satisfaga esta condición, puede ser considerado en mecanismo de DP.

Existen varios mecanismos que pueden cumplir con estas restricciones. Entre ellos, los dos que analizaremos en este apartado, el mecanismo laplaciano y el gaussiano. Como se presentó en la sección 1.3.4., ambos mecanismos lo que hacen es agregar ruido a la salida de una consulta, pero cumpliendo la garantía formal de DP. Asimismo, habíamos señalado que el mecanismo gaussiano tiende a funcionar mejor cuando se están realizando múltiples consultas secuenciales, ya que el orden en el que crece la complejidad de la sensibilidad es menor que para el mecanismo laplaciano. No obstante, el mecanismo gaussiano tiene ciertas desventajas. Primero, la distribución normal es mucho más acampanada que la laplaciana, por lo que a pesar de tomar ventaja de niveles de sensibilidad menores, no siempre va a brindar mejores resultados. En segundo lugar, utilizar el mecanismo gaussiano abre la puerta a eventos extremos de goteo de datos, problema del que no sufre el mecanismo laplaciano. Por estos motivos, aunque cada mecanismo presenta diferencias y ventajas teóricas para situaciones específicas, a continuación evaluaremos la *performance* de los mismos sobre los dos conjuntos de datos que tenemos, a fin de determinar cuál alcanza mejores resultados sobre cada conjunto de datos.

2.4.2. Tratamiento de inconsistencias en las tablas de distribución de frecuencias.

Un punto que no ha sido considerado aún es la posible inconsistencia de una tabla de distribución de frecuencias resultante de un mecanismo de DP. Como los mecanismos de DP agregan ruido tomado de variables aleatorias independientes e idénticamente distribuidas (IID) a cada punto de la salida, es decir, al conteo o probabilidad de cada categoría de una variable, es posible que su salida no sea consistente (Li *et. al.*, 2017). Veamos el siguiente ejemplo.

Gráfico VII. Distribución de probabilidades antes y después de aplicar DP para la variable ficticia 'Mascotas en los hogares'



Imaginemos que vamos a aplicar un mecanismo de DP, ya sea el laplaciano o el gaussiano sobre esta variable imaginaria llamada 'Mascotas en los hogares'. El gráfico anterior muestra la distribución de probabilidades para las categorías de esta variable ficticia, para la distribución original, y la nueva, una vez aplicada DP. Notemos dos cosas de importancia. Primero, la más obvia es que la probabilidad de la categoría 'Pez' quedó por debajo de 0 luego de aplicar el mecanismos de DP ¡Esto no tiene sentido! Segundo, la suma de las probabilidades del mecanismo de DP no suman 1, sino 0.905. Esto también representa un problema.

A pesar del desafío que suponen este tipo de inconsistencias, éstas pueden ser abordadas sin poner en riesgo las garantías formales de DP. Para ello nos valdremos de una deseable propiedad de los mecanismos de DP: sus resultados son inmunes al postprocesamiento. Es decir, una vez que se le aplicó el mecanismo de DP al conjunto de datos, podemos hacer cualquier tipo de transformación y análisis sobre los mismos, sin riesgo alguno de goteo de datos. Esto abre la posibilidad a que restauremos la consistencia interna de la distribución de frecuencias.

El problema más sencillo para abordar es la suma de probabilidades que difiere de 1. En ese escenario, lo que se puede hacer es normalizar la salida, para reponer los valores dentro del rango 0-1. En cambio, un problema que requiere más atención es el de las

categorías con probabilidad negativa. En ese caso, se pueden seguir varios enfoques. Una posibilidad es truncar el dominio de la distribución de la que el mecanismo toma el ruido. Así, podemos asegurarnos que los valores van a quedar dentro de cierto rango. Si algún valor cae por fuera de este rango, el mismo es reemplazado por el extremo del rango más cercano. Otra posibilidad es llevar todas las probabilidades negativas a 0, y distribuir (sustrayendo) proporcionalmente esta densidad de probabilidad de las categorías con probabilidad positiva. Nosotros hemos optado por este último enfoque, ya que distorsiona menos la distribución de probabilidades que el truncado arbitrario de los valores que sigue el método anterior. Hemos corrido ambas variantes de postprocesamiento 1000 veces sobre el conjunto de datos I, utilizando el mecanismo laplaciano y evaluando diferentes presupuestos, y los resultados fundamentan esta decisión. Los detalles de este procesamiento pueden seguirse en la entrada 2.2.1 de los notebooks de la tesis.

Grafico VIII. Error promedio del mecanismo laplace con postprocesamiento truncado y sin truncar sobre el dataset I. Resultados de 1000 ejecuciones.

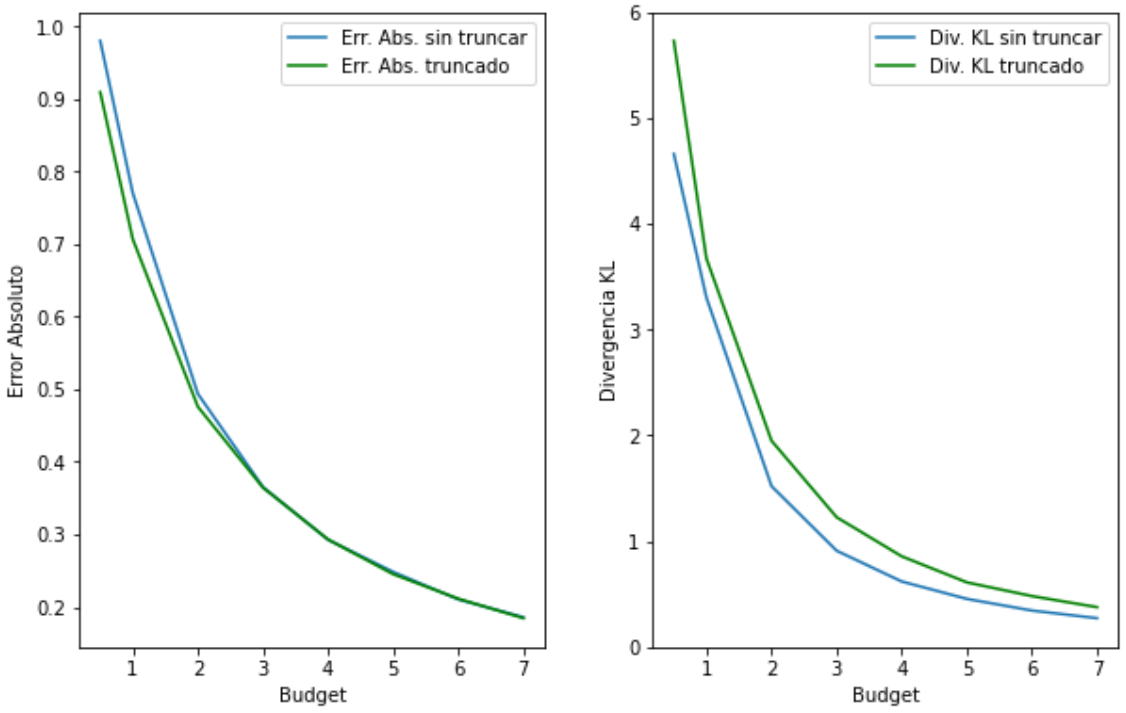


Tabla XIII. Error promedio del mecanismo laplace con postprocesamiento truncado y sin truncar sobre el dataset I. Resultados de 1000 ejecuciones.

Presupuesto	Postprocesamiento	Divergencia KL	Error Absoluto
0.5	Sin truncar	4.660449	0.979725
0.5	Truncado	5.730188	0.908970
1.0	Sin truncar	3.302976	0.770222
1.0	Truncado	3.672690	0.706637
2.0	Sin truncar	1.519092	0.493524
2.0	Truncado	1.948275	0.475852
3.0	Sin truncar	0.908394	0.365331
3.0	Truncado	1.225196	0.364405
4.0	Sin truncar	0.617325	0.292780
4.0	Truncado	0.855308	0.293010
5.0	Sin truncar	0.452802	0.248400
5.0	Truncado	0.609153	0.245241
6.0	Sin truncar	0.342353	0.210514
6.0	Truncado	0.479096	0.211485
7.0	Sin truncar	0.270195	0.186055
7.0	Truncado	0.374458	0.184675

Lo que señala la evidencia es que el postprocesamiento sin truncado presenta mejores resultados medidos a través de la divergencia KL y resultados similares, medidos a través del error absoluto, en relación al postprocesamiento con truncado. Estas conclusiones no son de sorprender, ya que en el post procesamiento es imposible alterar el error absoluto. Por eso, esta métrica nunca variará. No obstante, lo que sí cambia es en qué puntos de la distribución se agrupa ese error. En este sentido, truncar el rango de valores de los que se toma el ruido lo que hace es sobredimensionar la probabilidad de que una observación sea tomada de los extremos. La alternativa, de no truncar el rango, sino redistribuir proporcionalmente entre el resto de las categorías la densidad de probabilidad negativa, tiende a mantener mejor la estructura general de la distribución. Por ello, la divergencia KL muestra mejores resultados aplicando esta estrategia de post procesamiento. Para que el

lector pueda evaluar mejor el funcionamiento comparativo de ambas estrategias, en el anexo III de la tesis presentamos el pseudocódigo de las mismas.

2.4.3. Evaluación de mecanismos de *Differential Privacy* sobre el conjunto de datos I

Habiendo definido la estrategia conveniente para asegurar la consistencia de la tabla de distribución de frecuencias, pasaremos a evaluar los dos mecanismos que consideraremos, el laplaciano y el gaussiano, a fin de determinar cuál tiene mejor *performance* sobre el conjunto de datos I. Como se hizo anteriormente, evaluaremos los mecanismos de DP utilizando diferentes presupuestos. Asimismo, para garantizar la solidez de las conclusiones, correremos 1000 veces los mecanismos y reportaremos el error promedio. El detalle de la implementación de este ejercicio puede consultarse en la entrada 2.2.2 de los notebooks de la tesis.

Gráfico IX. Error promedio de los mecanismos de DP sobre el conjunto de datos I. Resultados de 1000 ejecuciones.

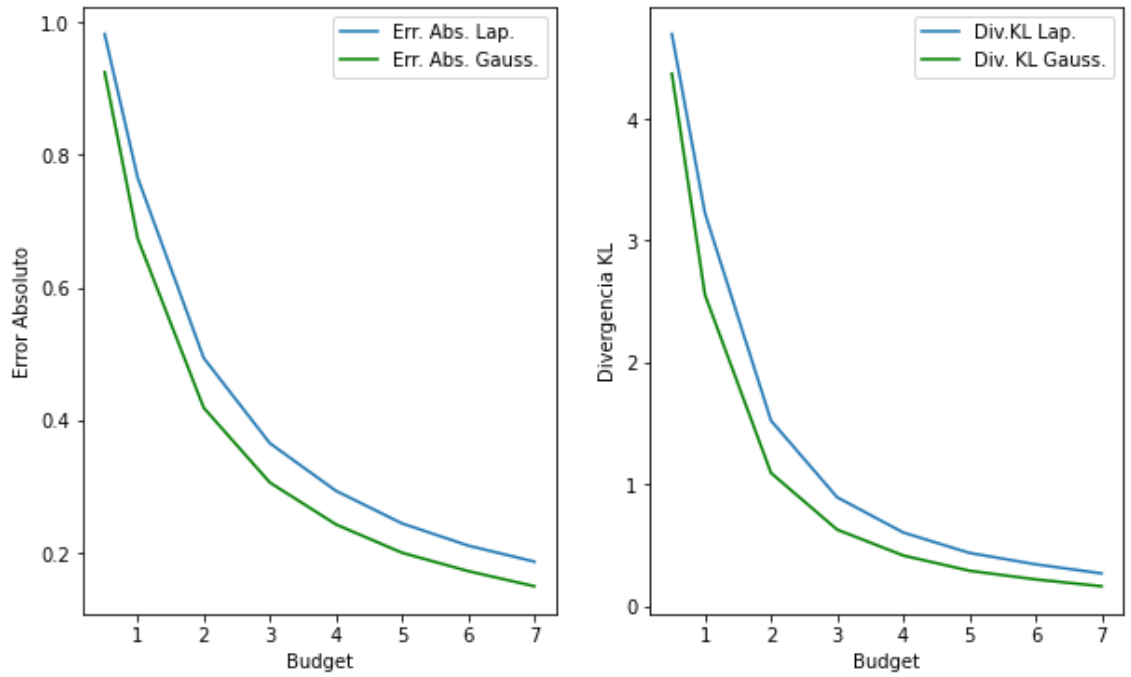


Tabla XIV. Error promedio de los mecanismos de DP sobre el conjunto de datos I. Resultados de 1000 ejecuciones.

presupuesto	Mecanismo	Divergencia KL	Error absoluto
0.5	Laplaciano	4.696936	0.981949
0.5	Gaussiano	4.371245	0.924574
1.0	Laplaciano	3.227135	0.765233
1.0	Gaussiano	2.556147	0.673612
2.0	Laplaciano	1.523102	0.493313
2.0	Gaussiano	1.093168	0.418248
3.0	Laplaciano	0.893661	0.364914
3.0	Gaussiano	0.628061	0.305717
4.0	Laplaciano	0.604455	0.293205
4.0	Gaussiano	0.415739	0.242548
5.0	Laplaciano	0.436947	0.244057
5.0	Gaussiano	0.290748	0.199844
6.0	Laplaciano	0.343549	0.210667
6.0	Gaussiano	0.219415	0.172202
7.0	Laplaciano	0.268976	0.186517
7.0	Gaussiano	0.163432	0.149430

Lo que evidencian estos resultados es que en esta comparación el mecanismo gaussiano tiene dominancia sobre el laplaciano. Para todos los presupuestos, la calidad de la salida de este mecanismo es superior al mecanismo laplaciano ¿A qué se debe esta mejor *performance*? A priori, la explicación más sencilla podría ser que el mecanismo gaussiano está obteniendo ventaja de utilizar la sensibilidad L2. Si bien esta es una ventaja, en realidad lo que esto significa en términos estrictos es que la sensibilidad, medida a través de la norma L2, crece en un orden inferior a la norma L1. Para k consultas correlacionadas, la norma L1 progresa en orden de k, mientras que la norma L2, progresa en orden $\sqrt{k}$. No obstante, la distribución gaussiana tiene la principal desventaja de que es más acampanada que la laplaciana, con lo que los valores tienden a estar más alejados del centro y mayor dispersión. Por ello, no debe darse por sentado que la utilización de uno u otro mecanismo

dará *de facto* mejores resultados. Esto debe evaluarse siempre en función de las consultas que vayan a realizarse.

Repasemos los parámetros de nuestras consultas y tratemos de comprender cómo están funcionando ambos mecanismos. Para el dataset en cuestión, tenemos 16 atributos, con lo que buscamos construir 16 tablas de distribución de frecuencias. Entonces, la cantidad máxima de consultas correlacionadas que estaremos realizando van a ser también 16. Asimismo, sabemos que la sensibilidad individual de cada una de estas consultas es de 1. Para todo este grupo de consultas, la sensibilidad medida a través de la norma L1 va a ser 16, y medida a través de la norma L2, va a ser 4. En base a estos parámetros, estudiemos las distribuciones de las que estará tomando el ruido cada mecanismo. Para ello, hagamos un breve repaso del funcionamiento de ambos mecanismos y especialmente de los parámetros utilizados para determinar las respectivas distribuciones de donde toman el ruido.

El mecanismo laplaciano puede formalizarse del siguiente modo:

$$F(X) = f(x) + Lap(\Delta_1 f/\varepsilon)$$

Mientras que el gaussiano puede definirse como:

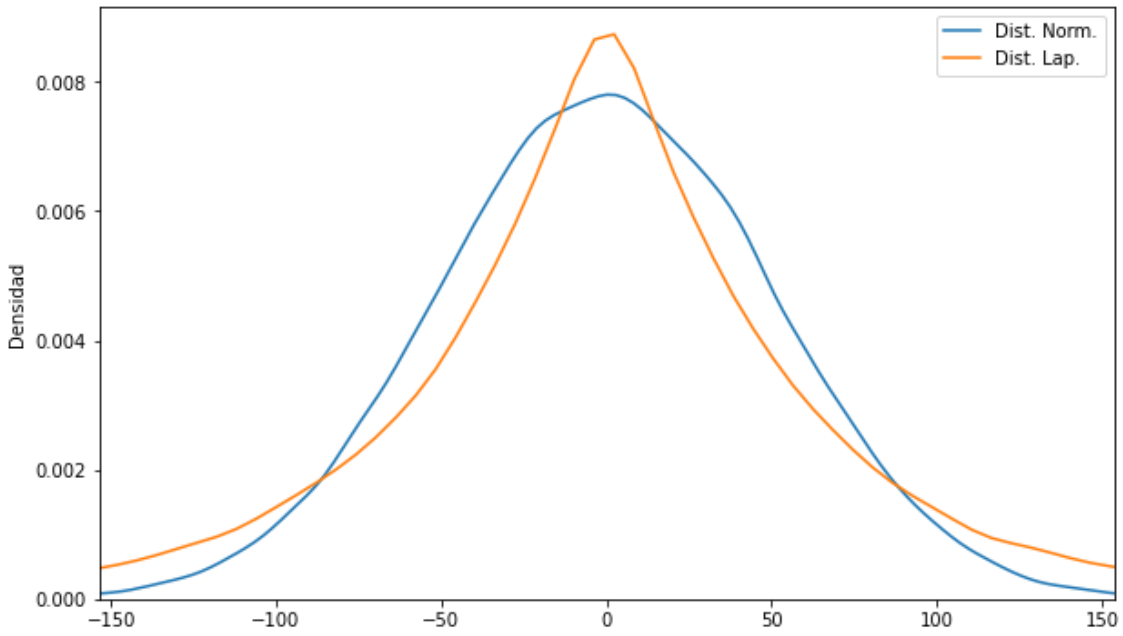
$$F(X) = f(x) + N\left(0, \sigma^2\right)$$

Donde:

$$\sigma^2 = \frac{2\Delta_2^2 \ln\left(\frac{1.25}{\delta}\right)}{\varepsilon^2}$$

Para construir ambas distribuciones, utilizaremos un presupuesto global de 5ε (a distribuir entre los 16 atributos) y un δ de 0,0005 (para el mecanismo gaussiano).

Gráfico X. Distribuciones Normal y Laplace de las que toman el ruido los mecanismos implementados.



Estas dos distribuciones se asemejan bastante. A pesar de que la distribución normal estándar es más “achatada” que la laplace estándar, en este caso ambas tienen una distribución de densidad muy parecida. Aunque la distribución laplace tiene mayor densidad acumulada en torno al centro, vemos que también tiene colas más grandes que la distribución normal. Dado que la densidad acumulada en los extremos de la laplace es bastante superior a la normal, cuyas colas caen más rápido, la probabilidad de extraer valores “extremos” es mayor utilizando el mecanismo laplaciano. Es por este motivo que para este conjunto de consultas, de elevada sensibilidad, el mecanismo gaussiano distorsiona menos los resultados. Básicamente, ello se debe a que tiene menos probabilidad de agregar ruido tomado de los extremos de la distribución. Pero no siempre este va a ser el caso. Si la sensibilidad de las consultas fuese baja, la densidad acumulada en torno al centro de la distribución laplace es muy superior a la gaussiana, con lo que más que contrarresta el efecto de las colas anchas.

2.4.4. Evaluación de mecanismos de *Differential Privacy* sobre el conjunto de datos II

Para el segundo sub conjunto de datos seguiremos el mismo procedimiento utilizado anteriormente. Aplicaremos los mecanismos laplaciano y gaussiano a fin de determinar cuál

tiene mejor *performance* sobre este conjunto de datos: correremos 1000 veces ambos mecanismos y reportaremos el error promedio, utilizando también distintos presupuestos[34].

Gráfico XI. Error promedio de los mecanismos de DP sobre el conjunto de datos II. Resultados de 1000 ejecuciones.

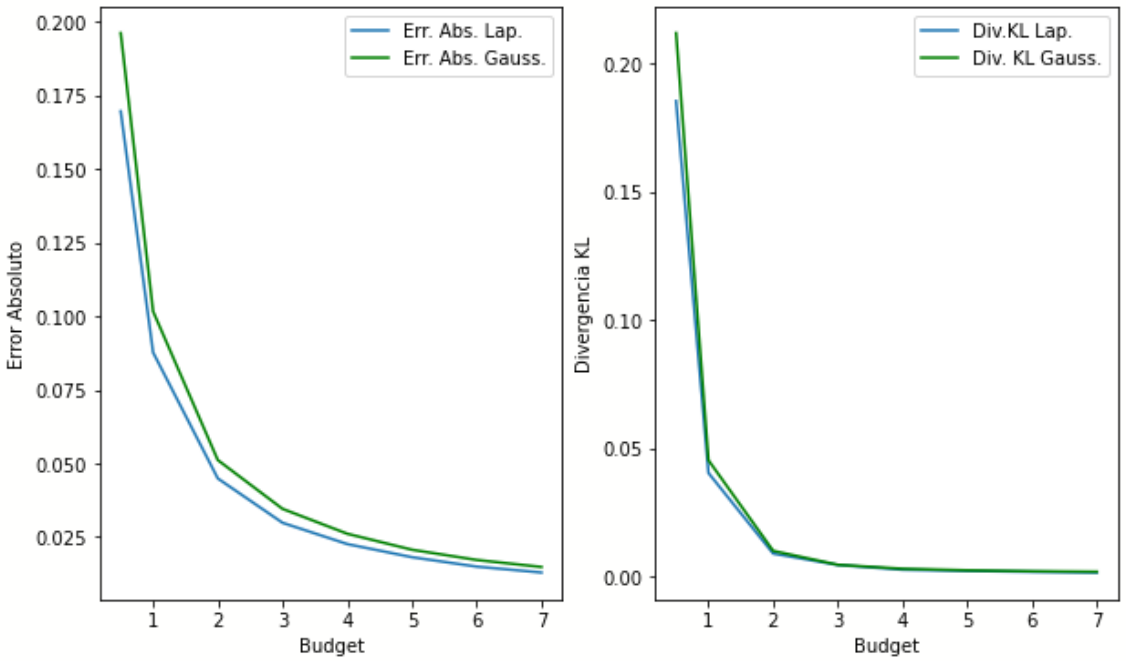


Tabla XV. Error promedio de los mecanismos de DP sobre el conjunto de datos II. Resultados de 1000 ejecuciones.

presupuesto	Mecanismo	Divergencia KL	Error absoluto
0.5	Laplaciano	0.185451	0.169673
0.5	Gaussiano	0.211951	0.196243
1.0	Laplaciano	0.040548	0.08772
1.0	Gaussiano	0.045508	0.101672
2.0	Laplaciano	0.00915	0.044879
2.0	Gaussiano	0.010206	0.051121
3.0	Laplaciano	0.00471	0.029813
3.0	Gaussiano	0.004779	0.03455
4.0	Laplaciano	0.002861	0.022583

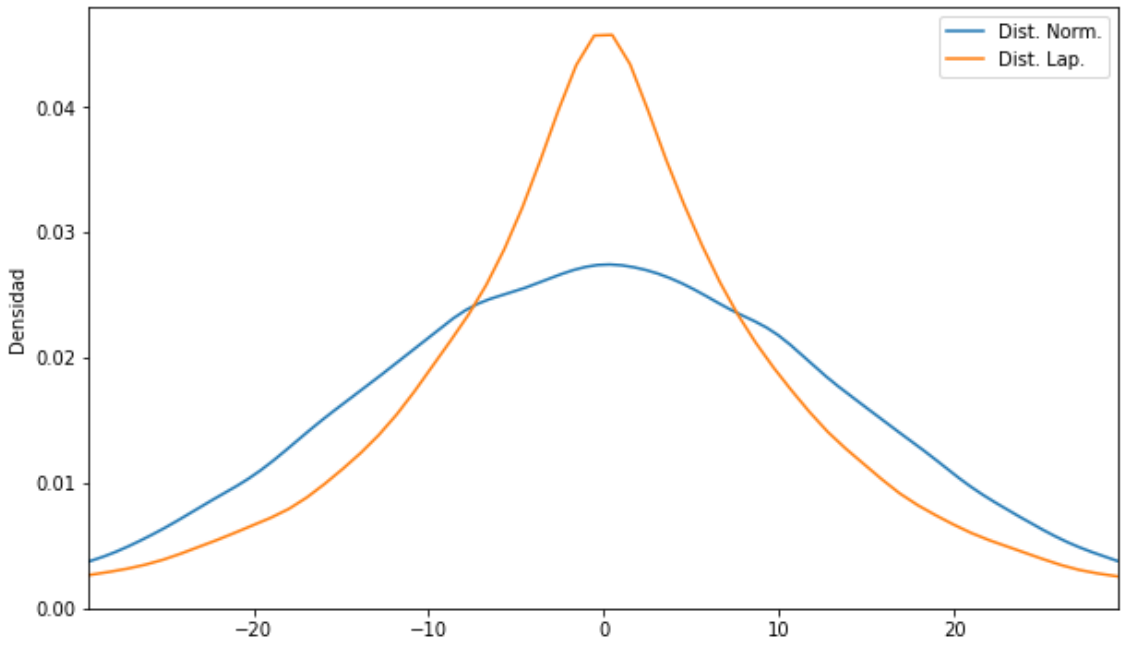
[34] El mecanismo gaussiano será implementado con un delta de 0.0005.

4.0	Gaussiano	0.003235	0.026089
5.0	Laplaciano	0.002406	0.018118
5.0	Gaussiano	0.002627	0.020653
6.0	Laplaciano	0.002046	0.014917
6.0	Gaussiano	0.002261	0.017178
7.0	Laplaciano	0.001772	0.012912
7.0	Gaussiano	0.002072	0.014799

Existen varios puntos interesantes en la comparación de estos resultados con los anteriores. Lo primero que resulta evidente es que las métricas de error para este dataset son muy inferiores a lo que eran sobre el primer dataset. Podríamos decir incluso que para este conjunto de datos, el error es virtualmente 0. Un punto a tener en cuenta para entender estos resultados es la sensibilidad de las consultas. Para el conjunto de datos II, la sensibilidad de las consultas es muy inferior a la del conjunto de datos I. Otro punto a considerar es el tamaño del dataset. El dataset II cuenta con casi cuatro veces más de registros que el dataset I, por lo que la adición de ruido tiene un impacto menor.

Pero el punto central a señalar de estos resultados es que a diferencia de lo que veíamos antes, ya no es el mecanismo gaussiano el que garantiza mejores resultados en nuestro experimento, sino el laplaciano. Para todo el conjunto de presupuestos evaluados, este mecanismo logró mejores resultados en este dataset en particular. Analicemos esta situación. En este caso, la sensibilidad medida por la norma L1 fue de 7, mientras que medida por la norma L2, la sensibilidad fue de $\sqrt{7}$. A diferencia del dataset anterior, donde la norma L1 era 4 veces superior a la norma L2, en este caso la brecha es menor. Entonces, la “ventaja” del mecanismo gaussiano se ve parcialmente neutralizada. Si bien tiene buena *performance* para publicar consultas de elevada sensibilidad, en consultas de sensibilidad más reducida el mecanismo gaussiano no funciona tan bien. Grafiquemos las dos distribuciones para seguir entendiendo lo sucedido.

Gráfico XII. Distribuciones Normal y Laplace de las que toman el ruido los mecanismos implementados.



En este caso la distribución laplaciana tiene una densidad cercana al centro muy superior a la de la normal. Esto quiere decir que utilizando esta distribución para muestrear el ruido, la probabilidad de tomar valores cercanos al centro es bastante mayor que utilizando una distribución normal. En cambio, la desventaja de la distribución laplaciana, que son las colas largas que acumulan una importante densidad en los extremos, no son tan relevantes en este caso, ya que acumulan mucha menos densidad que en el caso anterior. Por el contrario, la distribución normal, al ser más achatada, tiene mayor densidad acumulada en regiones alejadas del centro. Por ello, utilizando el mecanismo gaussiano la probabilidad de agregar valores de ruido elevados es mayor que si se utiliza el mecanismo laplaciano.

2.5. Definiendo los parámetros óptimos de los mecanismos de *Differential Privacy*

Hasta aquí hemos procesado los datos y definido qué tipo de mecanismo de DP conviene aplicar a cada conjunto de datos. En esta entrada buscaremos optimizar los parámetros de los mecanismos de DP para alcanzar los mejores resultados posibles. Asimismo, consideraremos diversos factores que pueden afectar a su *performance*.

2.5.1.Problemas de asignación del presupuesto entre atributos

Uno de los puntos a considerar para optimizar la *performance* de los mecanismos de DP es el criterio de asignación de presupuesto entre atributos. En los ejercicios anteriores, lo que se hizo fue repartir el presupuesto total en una magnitud similar entre todos los atributos. Ahora vamos a ver si existe alguna forma más eficiente de distribuir este presupuesto.

Durante el proceso de ingeniería de atributos uno de los objetivos planteados fue el de reducir la dimensionalidad de los datasets y la cardinalidad de los atributos. Conceptualmente, estas transformaciones apuntaron por un lado a disminuir la sensibilidad de las consultas, pero sobre todo, a disminuir la cantidad de categorías con un bajo número de observaciones, ya que son muy castigadas por la adición del ruido. Llegado este punto, una posibilidad para seguir mejorando la calidad de la salida, es redistribuir el presupuesto entre los atributos de tal modo que las variables de mayor cardinalidad reciban mayor presupuesto. De este modo, se lograría disminuir la cantidad de ruido que se agrega en estas variables, mientras que se incrementaría el ruido que se agrega en variables con menos categorías. Esta reasignación del presupuesto, podría dar lugar a una mejora en la calidad de la información.

Para evaluar esta hipótesis, hemos corrido sobre ambos conjuntos de datos por separado los mecanismos de DP seleccionados anteriormente. Primero los corrimos utilizando una asignación similar de presupuesto entre los atributos y luego repetimos el ejercicio utilizando una asignación de presupuesto en función de la cardinalidad de la variable. Finalmente comparamos los resultados tanto a nivel agregado, como abiertos por categorías. Estos experimentos fueron realizados con un presupuesto global de 4 ε y 0,0005 δ (sólo para el mecanismo gaussiano utilizado en el set de datos I). Para garantizar la robustez de los resultados todos los mecanismos fueron evaluados 1000 veces y el error reportado corresponde a un promedio del error del mismos. Para un mayor nivel de detalle del

procedimiento y de los resultados, remitimos al lector a las entradas de los notebooks de la tesis 2.2.3 y 3.2.2.

La siguiente tabla muestra el error de cada estrategia en la asignación del presupuesto para el experimento realizado en el set de datos I.

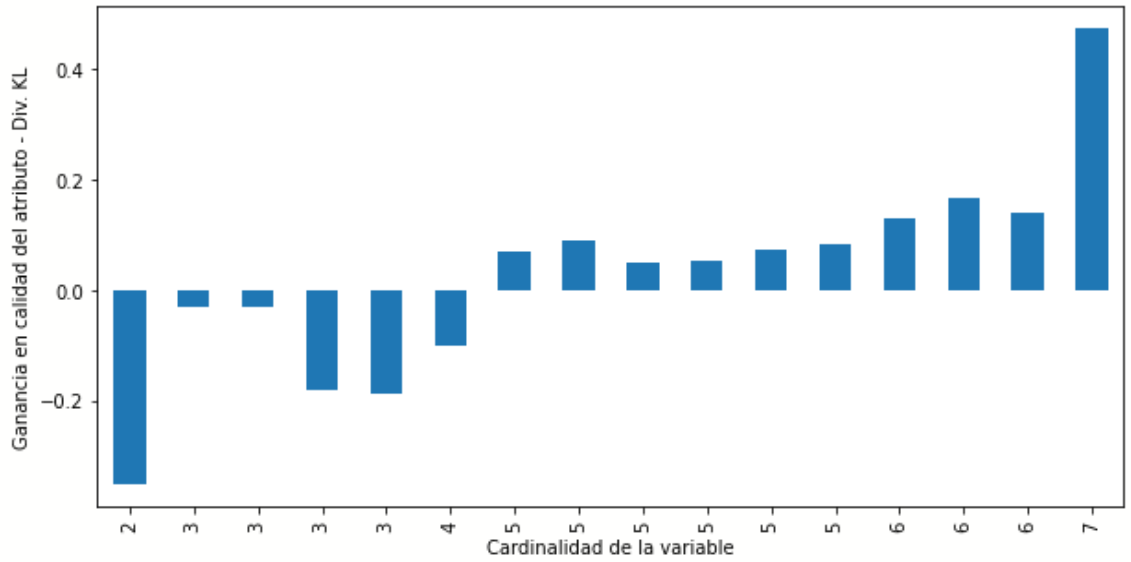
Tabla XVI. Comparación de mecanismos de asignación de presupuesto entre atributos. Métricas de error para el set de datos I.

Asignación del presupuesto	Divergencia KL	Error Absoluto
Similar	0.4090	0.2424
Por cardinalidad	0.3860	0.2420

El experimento señala que aunque el error absoluto no tiene mejoras importantes, la divergencia KL registra ciertas mejorías. Siguiendo la nueva estrategia de asignación del presupuesto la divergencia KL parece ser inferior a la anterior. Para a verificar la existencia de una diferencia significativa en términos estadísticos, hemos realizado un test de diferencias de medias sobre la divergencia KL. Para ello, hemos comparado la distribución de la divergencia KL de las 1000 ejecuciones con el criterio I de asignación del presupuesto, contra la divergencia KL de las 1000 ejecuciones con el otro criterio. Como no conocemos las distribuciones que siguen estas variables, hemos optado por realizar un *Wilcoxon signed-rank test* que es un tipo de test de diferencias de medias no paramétrico para muestras no independientes. Como ambas muestras fueron construidas aplicando mecanismos de DP sobre el mismo dataset, las mismas deben ser tratadas como muestras relacionadas. El resultado de este test permite rechazar la hipótesis nula de que ambas muestran provienen de la misma población. El p-valor para este test es 0.000046, por lo que rechazamos H0 con una significancia estadística del 1%.

Veamos ahora cómo se distribuyó esta ganancia en la calidad de la información de acuerdo a la cardinalidad de las variables.

Gráfico XIII. Ganancia en la calidad de la información según cardinalidad de la variable - Set de datos I.



Esta apertura de la mejora en la calidad de la información (medida a través de la divergencia KL), muestra exactamente los resultados que esperábamos: las variables de mayor cardinalidad resultaron beneficiadas, mientras que las de menor cardinalidad perdieron calidad. Esta reasignación del presupuesto trajo aparejada una mejora en la calidad global de la información.

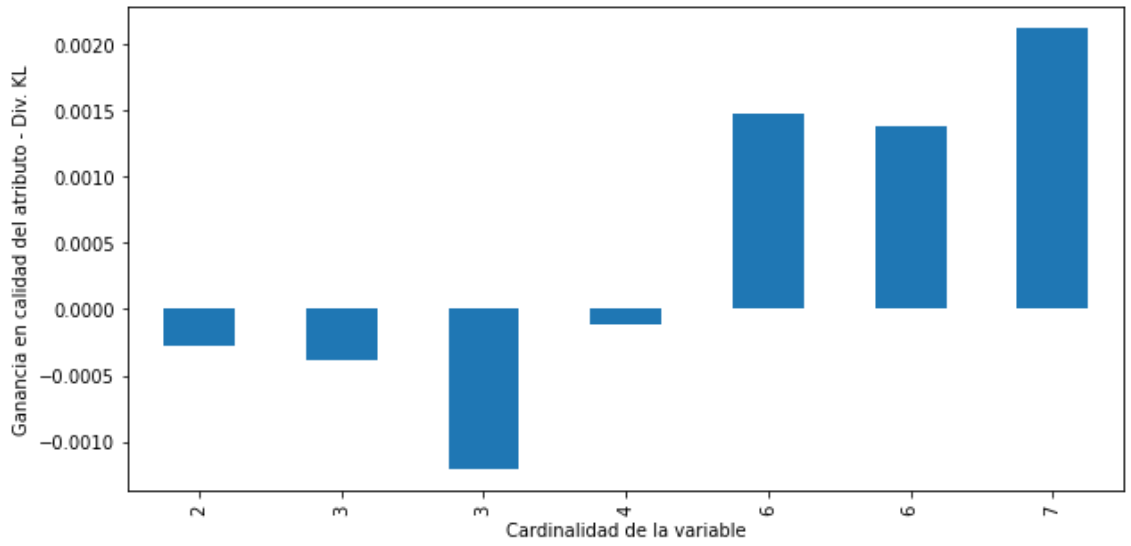
Tabla XVII. Comparación de mecanismos de asignación de presupuesto entre atributos. Métricas de error para el set de datos II.

Asignación del presupuesto	Divergencia KL	Error Absoluto
Similar	0.0033	0.0226
Por cardinalidad	0.0026	0.0219

Los resultados de la evaluación de estrategias de asignación de presupuesto sobre el set de datos II llevan a las mismas conclusiones presentadas más arriba. A pesar de que el mecanismo utilizado es diferente (en este caso utilizamos el mecanismo laplaciano en lugar del gaussiano) y de que el error se halla muy cercano a 0, la estrategia de asignación del presupuesto en función de la cardinalidad de las variables permitió una ganancia en la calidad de la información. Al igual que en el caso anterior, hemos realizado un *Wilcoxon signed-rank test* para determinar si esta diferencia es significativa en términos estadísticos.

En este caso, el p-valor del test de *Wilcoxon* fue 0.001763, por lo que podemos rechazar H0 (que ambas muestras provienen de la misma población) con una significancia estadística del 1%.

Gráfico XIV. Ganancia en la calidad de la información según cardinalidad de la variable - Set de datos II.



Vemos que una vez más se repiten los resultados anteriores. Las variables de alta cardinalidad han salido favorecidas con esta nueva asignación del presupuesto, mientras que las de baja cardinalidad han deteriorado la calidad de su información. Aunque su análisis excede los fines del presente trabajo, creemos que este esta estrategia de asignación del presupuesto es un buen principio con capacidad de generalización a otros conjuntos de datos, al margen de su estructura o del mecanismo de DP que esté siendo utilizado. De cara a futuras iteraciones, pueden probarse variantes de estos criterios de asignación del presupuesto. En este caso, la hipótesis fue que las variables de mayor cardinalidad tenían menos observaciones en cada categoría y consecuentemente sufrían más la adición de ruido. No obstante, aunque esto es lo más normal, también puede darse la posibilidad de que una variable con pocas categorías tenga baja entropía y consecuentemente alguna de sus categorías quede con pocas observaciones. Teniendo esto en cuenta, podría evaluarse incluir en un criterio de asignación de presupuesto no sólo la cardinalidad de la variable, sino también la entropía.

2.5.2. Optimización del presupuesto

2.5.2.1. Composición simple y avanzada

Habiendo definido el mecanismo a utilizar y el criterio de asignación del presupuesto entre atributos, es momento de hacer un análisis en profundidad del presupuesto de privacidad que usaremos. Aunque hasta ahora estuvimos incluyendo el presupuesto dentro de los análisis de sensibilidad realizados, no hemos discutido en detalle cuál es el presupuesto de privacidad óptimo para cada conjunto de datos. Por otro lado, tampoco hemos considerado la posibilidad de medir el presupuesto a través del teorema de composición avanzada, sino que lo hemos hecho únicamente a través de la composición simple.

El presupuesto puede ser computado siguiendo dos criterios: el de composición simple, y el de composición avanzada. La composición simple es bastante sencilla. Según el teorema de la composición simple, el presupuesto total de una una serie de consultas k, donde cada una es $(\varepsilon, 0)$- *differentially private* viene dado por:

$$\varepsilon' = k\varepsilon$$

En cambio, utilizando el teorema de la composición avanzada, el presupuesto total viene dado por:

$$\varepsilon' = \sqrt{2k\ln(1/\delta'')}\epsilon + k\epsilon(e^{\varepsilon} - 1)$$

Donde δ'' representa un slack que se adiciona al presupuesto. Este slack, al igual que el parámetro δ' del mecanismo de DP (que para este ejemplo es 0), representa una probabilidad de goteo de datos. Al aceptar este riesgo adicional, se puede hacer uso del teorema de la composición avanzada. Este teorema lo que nos permite es, a cambio de asumir cierto riesgo extra representado por el slack, realizar las mismas consultas que realizamos en el escenario de composición simple, pero a un costo menor. Esto brinda grandes ventajas a la hora de trabajar con una gran cantidad de consultas, ya que permite reducir el presupuesto global o, lo que es lo mismo, a similar presupuesto, aumentar el nivel de privacidad. Bajo el principio de composición simple, tenemos que el presupuesto global ε' se halla en el orden de $k\epsilon$, mientras que bajo composición avanzada, tenemos que ε' se halla en el orden de $k\epsilon^2 + \sqrt{k}\epsilon$.

No obstante, a pesar de las ventajas que emanan de este teorema, debemos tener en cuenta que ellas dependan de los parámetros ε' y δ''. Adicionalmente, también debemos considerar que si trabajamos con el mecanismo gaussiano (que es lo que en general vamos a utilizar si tenemos una gran cantidad de consultas), este slack δ'' va a estar agregándose al δ' propio del mecanismo, con lo que el riesgo de un evento adverso de goteo de datos (si los parámetros no son bajos), comienza a acumularse peligrosamente. Por estos motivos, y a pesar de las ventajas teóricas de la composición avanzada, debemos evaluar su conveniencia sobre el caso de nuestro interés.

2.5.2.2. Análisis de sensibilidad del error del mecanismo de Differential Privacy en función del presupuesto. Set de datos I.

A continuación haremos el análisis de sensibilidad del error en función del presupuesto computado tanto mediante el principio de composición simple, como de composición avanzada. Utilizaremos un δ de 0.0005 para el mecanismo de DP y permitiremos un slack de 0.005 para computar el presupuesto bajo composición avanzada. Repetiremos los cálculos 100 veces para asegurar la solidez de los resultados y reportaremos las métricas como el error promedio de todas las ejecuciones. El detalle de este procesamiento puede consultarse en la entrada 2.2.4 de los notebooks de la tesis.

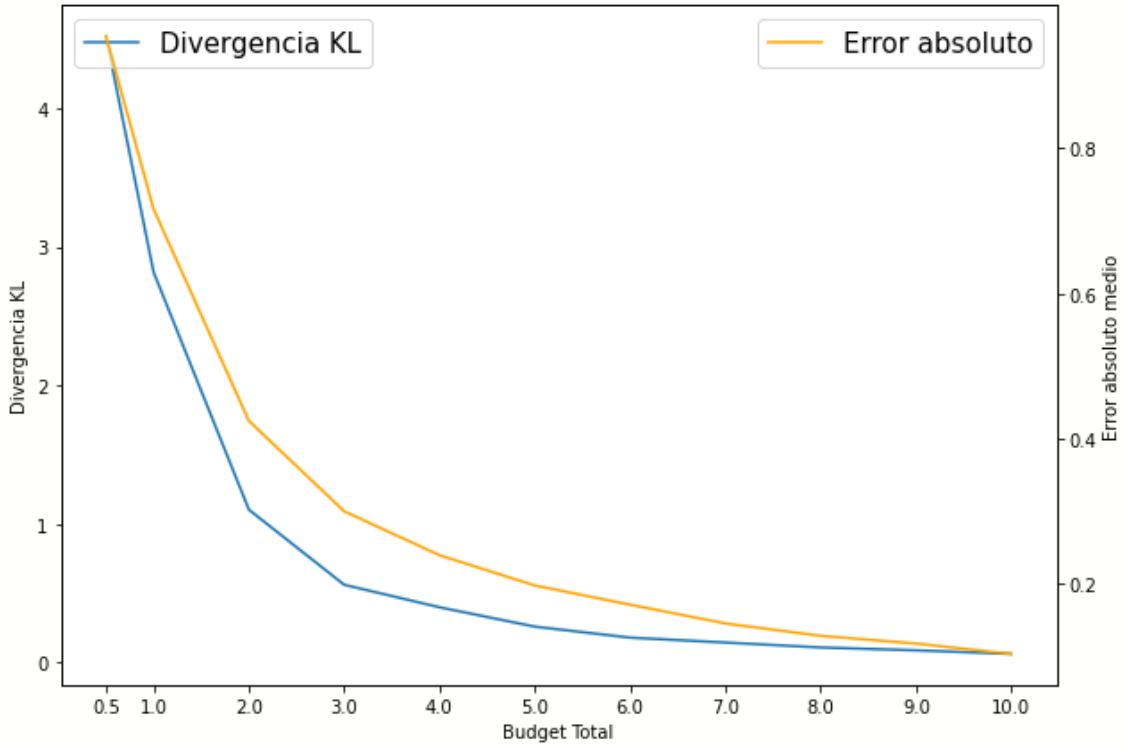
Tabla XVIII. Análisis de sensibilidad del error promedio del mecanismo de DP en función del presupuesto para el set de datos I. Resultados de 100 ejecuciones.

Presupuesto Composición Simple	Presupuesto Composición Avanzada	Delta	Div. KL	Error Absoluto
0.5	0.34685	0.01293	4.52066	0.954201
1.0	0.772783	0.01293	2.81677	0.716117
2.0	1.731923	0.01293	1.102445	0.424736
3.0	2.798768	0.01293	0.56029	0.30025
4.0	3.939388	0.01293	0.398058	0.23984
5.0	5	0.01293	0.258123	0.198215
6.0	6	0.01293	0.178785	0.171723
7.0	7	0.01293	0.143408	0.14595

8.0	8	0.01293	0.107792	0.128863
9.0	9	0.01293	0.087129	0.118301
10.0	10	0.01293	0.061471	0.104061

El primer elemento a remarcar de estos resultados es que para niveles de error tolerables, en la zona del error absoluto inferior a 0.30, la diferencia entre el presupuesto computado como composición simple y avanzada es inexistente. Incluso más, computado como composición avanzada, estamos aceptando un δ' total muy elevado (sin obtener ventaja alguna). En este caso, la composición avanzada no representa beneficios. Ello se debe a que la composición avanzada tiende a mejorar el presupuesto global a medida que se incrementan la cantidad de consultas, o que se reduce el presupuesto de cada consulta individual. En la zona donde nos estamos moviendo nosotros, no reporta ventajas significativas. En lo que sigue de las mediciones, nos manejaremos únicamente utilizando la composición simple.

Gráfico XIX. Sensibilidad promedio del error en función del presupuesto (composición simple) - Set de datos I. Resultados de 100 ejecuciones.



Para todo el dominio de este gráfico el δ' total insumido por el mecanismo gaussiano es de 0.008 (que surge de un δ de 0.0005 para cada consulta). En lo que refiere al análisis de sensibilidad realizado en función del presupuesto, lo que vemos es que entre los 2 y 4 ε' la

divergencia KL acumula una caída superior al 60%, pero ya en adelante este descenso se suaviza. A priori, la revisión visual de este gráfico sugiere que en torno a los 3 o 4 ε' se alcanza un 'codo' en la caída del error, lo que señala un punto interesante para ubicar los valores de presupuesto óptimos. No obstante, debemos evaluar si esta región ofrece un equilibrio aceptable en términos del error (que puede ser muy alto como para aceptarse), como de privacidad (la garantía puede ser muy débil). Asimismo, no debemos dejar fuera de consideración la posibilidad de inexistencia de un punto óptimo aceptable bajo los criterios de privacidad y calidad de la información.

Dado el rango establecido entre 3 y 4 ε', notamos que las métricas de error para el primer extremo son muy elevadas. En efecto, un nivel de error absoluto de 0,30 parece muy elevado como para poder aceptarlo. Por ello, optaremos por seguir trabajando con un presupuesto de $4\varepsilon'$, donde el error absoluto disminuye a 0,23. No obstante, este presupuesto no es ideal en términos de privacidad. Se considera que un presupuesto menor a $1\varepsilon'$ garantiza fuertes niveles de privacidad. En cambio, a medida que nos alejamos, de 1, la pérdida de privacidad progresa muy rápidamente. El presupuesto establecido de $4\varepsilon'$ implica niveles **bajos** de privacidad, niveles que probablemente en la mayoría de los casos no sean aceptables, pero que nosotros podemos tolerar. Sobre este punto volveremos más adelante.

Dado que el set de datos II ha reportado consistentemente métricas de error muy inferiores a las del set de datos I, no realizaremos este análisis sobre el mismo, y tomaremos el presupuesto que se determine para este conjunto de datos como válido para aquel también.

2.5.2.3. Consideraciones sobre los riesgos a la privacidad

Dado que estamos aceptando trabajar en un entorno de privacidad laxa, debemos justificar esta decisión. Recordemos de la presentación del problema que nosotros estamos abordando **solo uno de los medios a través de los cuales el juzgado hace pública la información relativa a las causas**. Existen dos mecanismos a través de los cuales la información contenida en las resoluciones se distribuye: la publicación on-line de los textos completos de las resoluciones (previa aplicación de técnicas tradicionales de anonimización) y la publicación de un set de datos en formato tabular, que es el medio sobre el cual estamos trabajando nosotros. Por mandato legal y profesional, la publicación del texto escrito de las resoluciones es un requisito fundamental. El juzgado no puede discontinuar la difusión a través de este medio, ni aplicar técnicas de anonimización avanzadas que puedan

afectar en exceso al texto original. Esto implica que siempre va a quedar una puerta abierta al acceso a los datos originales de las causas. Por este motivo, no tiene sentido establecer altos estándares de privacidad en lo que hace a la publicación del set de datos tabulares, ya que la posibilidad de amenazas a la privacidad nunca va a conjurarse completamente.

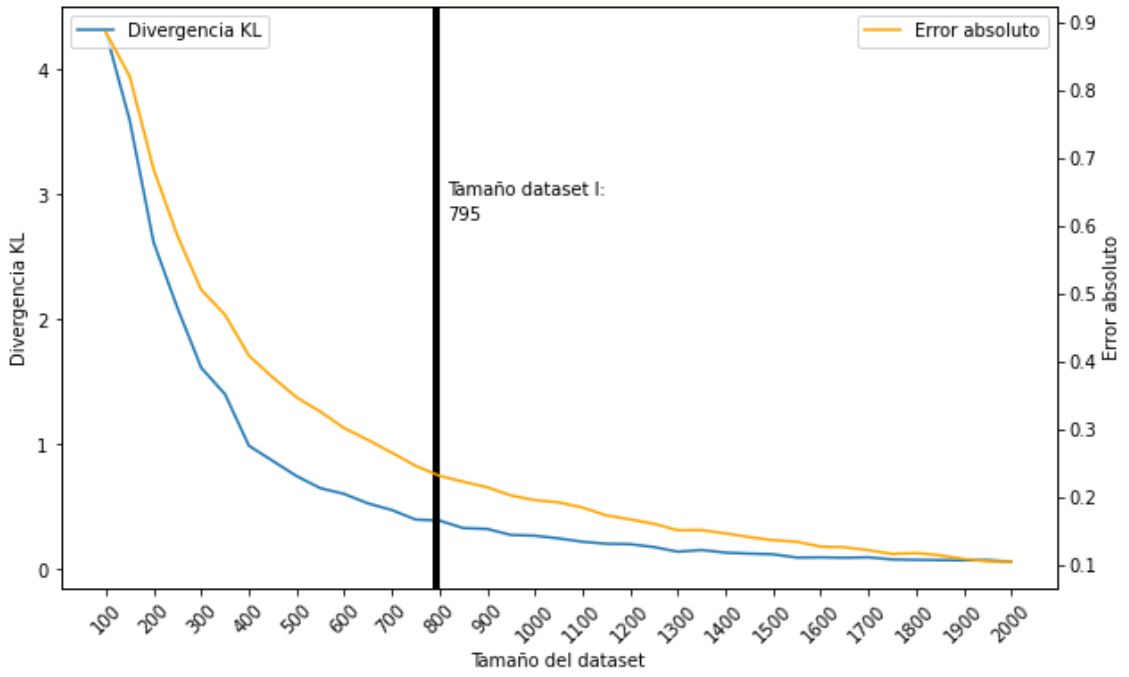
Dado este contexto ¿Tiene sentido proponer el uso de DP si sabemos que el riesgo de goteo de datos siempre está latente? Nosotros creemos que sí. En primer lugar, el ejercicio de esta tesis se limita únicamente a dar garantías formales sobre el set de datos que estamos abordando. Pero, pensando en términos más amplios, creemos que todo esfuerzo que conlleve a una mejora en la privacidad de los involucrados, es satisfactorio. Naturalmente, el juzgado no va a poder dar garantías formales sobre la cantidad de información que un atacante puede ganar, pero en cambio puede asegurar que dentro del marco normativo en el que se mueve, está aplicando los mejores estándares disponibles. Por otro lado, debe tenerse en cuenta que los involucrados en las causas no tienen expectativa de privacidad sobre su participación en las causas, toda vez que los procesos pueden ser públicos, e incluso en los casos donde no lo sean, las resoluciones sí lo son. Por ello, creemos que protegiendo la privacidad en el contexto de la publicación del dataset estructurado (aunque la misma información pueda consultarse en el texto escrito de las resoluciones), estamos defendiendo la expectativa de privacidad de los individuos. Todo esto no implica que el flujo de trabajo propuesto no deba mejorarse, al contrario, el ejercicio en cuestión es solo el primer paso en el desarrollo de un flujo de trabajo que permita conjugar satisfactoriamente la necesidad de difusión de la información, con las demandas de privacidad de los involucrados.

2.5.3. Sensibilidad del error al tamaño del dataset

Aunque no es un parámetro del modelo que podamos controlar, debemos analizar cuál es la relación entre el tamaño del dataset y la calidad de la información para realizar recomendaciones de cara a futuras mejoras. Sabemos que a menor cantidad de observaciones disponibles para cada consulta, menor es la calidad de la salida debido a la adición de ruido. Pero ¿La cantidad de registros de la que disponemos es al menos suficiente para garantizar una calidad aceptable en la salida? Para responder esta pregunta hemos generado datasets bootstrapeados de diverso tamaño y hemos evaluado el error en la salida luego de aplicar DP en base a los parámetros antes definidos. Al igual que en los casos anteriores, hemos repetido las estimaciones 100 veces y reportamos el error como un

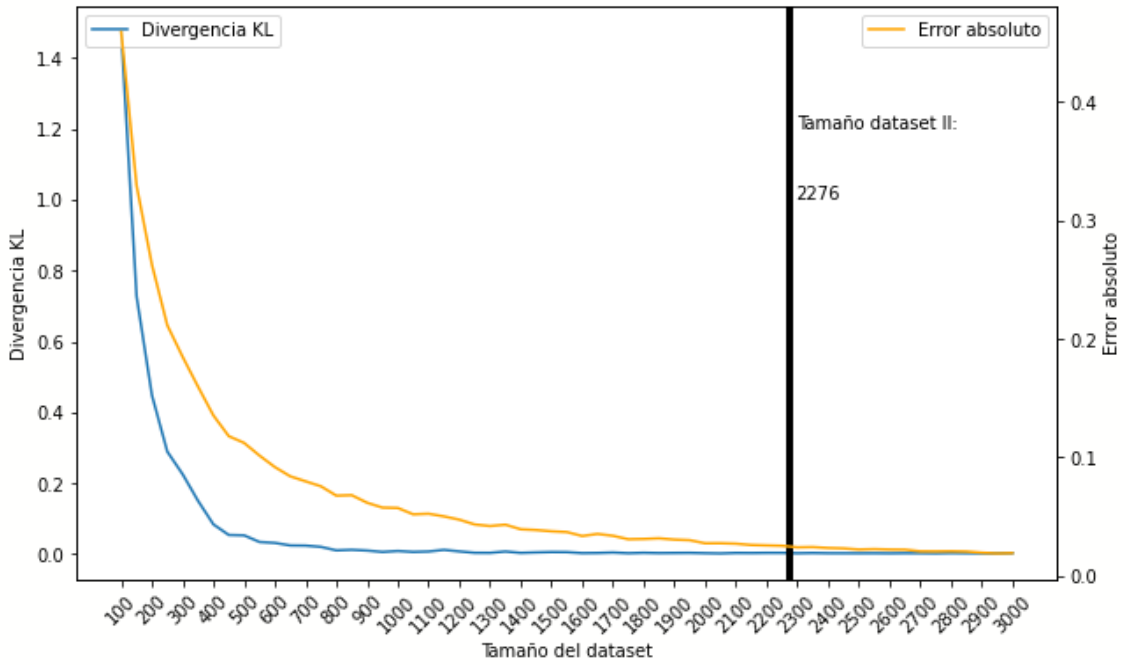
promedio del error de todas las ejecuciones. Estos ejercicios pueden seguirse en las entradas 2.2.5 y 3.2.3 de los notebooks de la tesis.

Gráfico XX. Sensibilidad promedio del error en función del tamaño del dataset - Set de datos I. Resultados de 100 ejecuciones.



En base a nuestras ejecuciones, podemos señalar que la cantidad de observaciones con las que cuenta el dataset I es bastante reducida y se halla lejos de ser óptima. De hecho, duplicando el tamaño del dataset el error medido a través de la divergencia KL se reduciría en más del 70%. Y de triplicar el número de observaciones, la divergencia KL sería alrededor de un 85% inferior y el error absoluto, un 60%. Esta comprobación pone de manifiesto una vez más que los mecanismos de DP son 'hambrientos' de datos. Con un número mayor de observaciones, no sólo podríamos disminuir el error de la salida, sino utilizar un presupuesto más bajo, con lo que la garantía de privacidad sería más estricta. Asimismo, de contar con un gran número de observaciones, podría publicarse más volumen de información (más atributos), e incluso se podrían construir tablas de contingencia, que son incluso más demandantes de observaciones que las tablas de distribución de frecuencias.

Gráfico XXI. Sensibilidad promedio del error en función del tamaño del dataset - Set de datos II. Resultados de 100 ejecuciones.



Para el set de datos II, al contar con mayor cantidad de observaciones, las métricas de error no tienen tanto margen para disminuir. De hecho, la divergencia KL para la cantidad de observaciones disponibles es virtualmente 0 y el error absoluto está en torno al 2,5%. No obstante, lo mismo que dijimos anteriormente vale para este set de datos. Con una mayor cantidad de observaciones podrían pensarse nuevas (y mejores) estrategias para publicar la información, como publicar nuevos atributos, tablas de contingencia e incluso ofrecer garantías más estrictas de seguridad. De todos modos, en lo que refiere a este conjunto de datos, y a la estrategia elegida actualmente, mayor cantidad de observaciones no redundaría en un beneficio significativo.

.

2.6. Recapitulando sobre la propuesta - Recomendaciones de cara a un *MVP.*

A lo largo de las secciones anteriores hemos probado que es posible implementar una estrategia de difusión de los datos alternativa a la publicación de datasets estructurados actualmente utilizada por el Juzgado. A pesar de los obstáculos y desafíos (principalmente vinculados a la reducida cantidad de observaciones), hemos logrado presentar una propuesta de trabajo alternativa, que mantiene a la vez la calidad de la información y ofrece mejores garantías de privacidad. A continuación presentamos las consideraciones de mayor relevancia a la hora de poner en práctica la propuesta. Asimismo, en el esquema IV (en la página 114) se presenta un flujo de trabajo que recupera los principales pasos a tener en cuenta, tanto en la etapa de definición del problema, como en la etapa de procesamiento de los datos.

Para la implementación de nuestra solución, nos hemos valido de los aportes de DP. En lugar de proponer estrategias de publicación complejas, como son la publicación de tablas de contingencia o datasets sintéticos, hemos optado por proponer la construcción y publicación de tablas de distribución de frecuencias. Por conveniencia en el procesamiento y para evitar posibles confusiones de parte del público usuario de la información, recomendamos presentar la información normalizada, no como una frecuencia absoluta, sino relativa.

En la etapa exploratoria de los datos hemos realizado una importante jerarquización y descarte de atributos que, o bien no eran de interés en función de los objetivos propuestos, o que siéndolo la calidad de la información era baja. En este sentido, descartamos atributos vinculados a aspectos administrativos de las causas, de apelaciones, o atributos con baja varianza o gran número de valores faltantes.

En lo que refiere al tratamiento de los datos, sugerimos dividir el conjunto en datasets acorde a la temática de las causas. Dado que las causas de violencia de género contienen mucha información específica, es conveniente separarlas del resto de las causas. Ello permite ajustar la estrategia de tratamiento de los atributos de forma más conveniente en base a las características de cada tipo de causas. En este sentido, un punto a considerar de cara a futuras iteraciones es el de mejorar los criterios de exposición de la información en lo que refiere al conjunto de causas diversas, ya que los atributos en general tienen gran cardinalidad y entropía, con lo que la calidad de la información así presentada es baja.

Posteriormente, hemos realizado una serie de propuestas en lo que refiere a las transformaciones de los atributos. En esta línea el objetivo en general fue el de reducir la dimensionalidad de los datasets y la cardinalidad de los atributos. Así, hemos fusionado atributos donde fue posible y reordenado la cardinalidad de las variables donde era elevada. Todo ello apuntó a disminuir el nivel de ruido incorporado por el mecanismo de DP. Vimos que las ventajas de estas transformaciones fueron bastante significativas para ambos conjuntos de datos.

Luego, hemos evaluado estrategias de postprocesamiento para garantizar la consistencia de las tablas de distribución de frecuencia resultantes de aplicar los mecanismos de DP. Hemos comparado la posibilidad de truncar las probabilidades en 0 (transformando todas las probabilidades negativas en 0) y también la alternativa de redistribuir la densidad de probabilidad negativa sobre el resto de la distribución en zona positiva. Hemos comprobado que esta última alternativa brinda resultados de mejor calidad que la anterior.

Siguiendo con los criterios de optimización de los mecanismos de DP, hemos considerado dos estrategias para la asignación del presupuesto entre los atributos. Por un lado, consideramos la posibilidad de asignar el mismo presupuesto a cada atributo. Por otro lado, consideramos la asignación del presupuesto de acuerdo a la cardinalidad de las variables, a fin de disminuir el ruido a adicionar en categorías con bajos conteos. Los resultados de la experimentación sobre ambos conjuntos de datos sugieren que este último criterio de asignación ofrece mejoras sensibles sobre la calidad de la información.

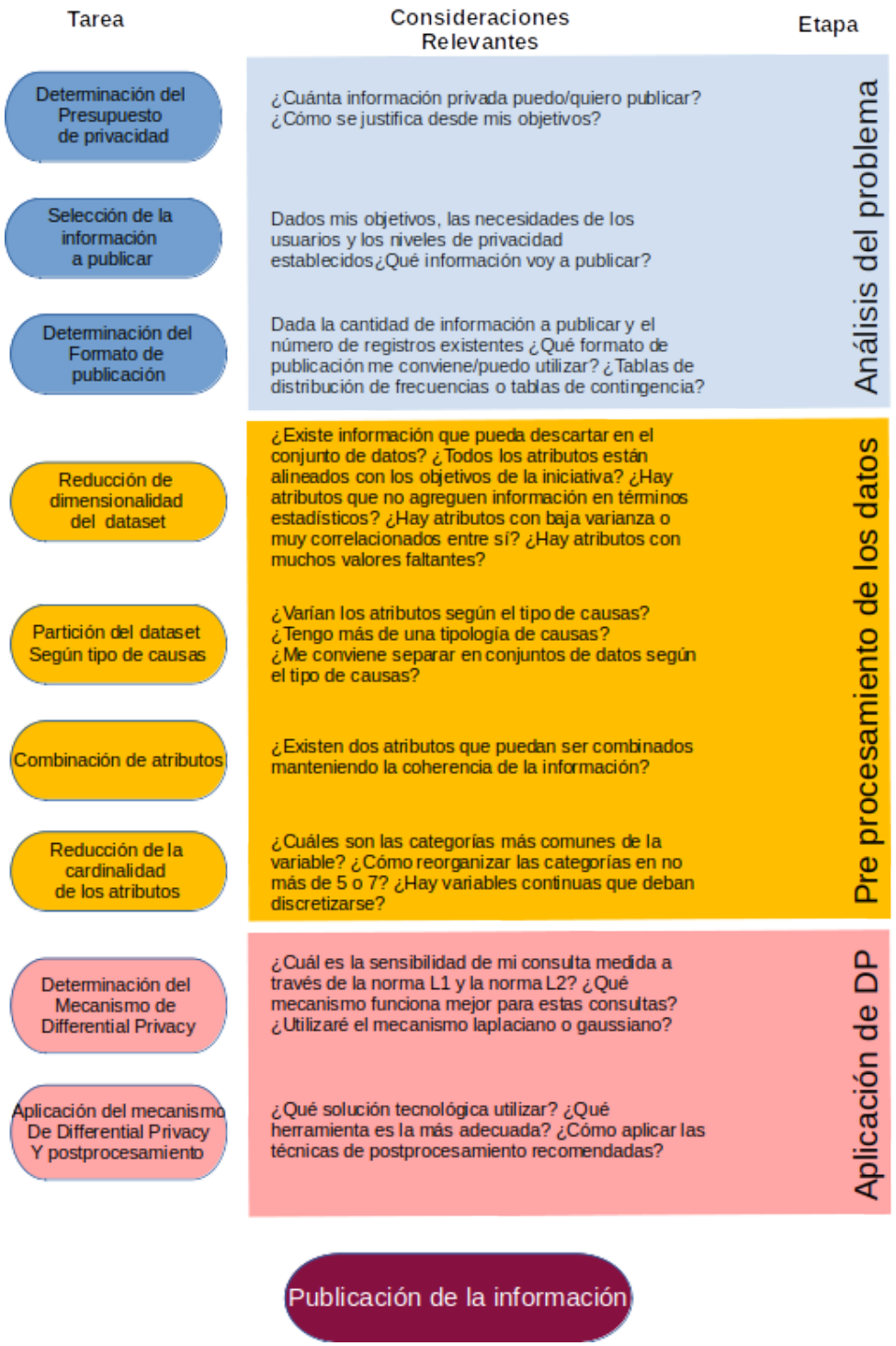
Para la determinación del mecanismo de DP a utilizar comparamos dos variantes, el mecanismo gaussiano y el laplaciano. Nuestras pruebas señalan que el mecanismo gaussiano funcionó mejor para el set de datos I que contiene las causas de violencia de género y que el mecanismo laplaciano funcionó mejor para el set de datos II, que contiene el resto de las causas. Básicamente, esto responde a la sensibilidad de las consultas realizadas sobre cada set de datos. Mientras que el mecanismo laplaciano es recomendable para consultas de sensibilidad baja, el mecanismo gaussiano tiene mejores resultados en contextos de sensibilidad elevada.

En base al desarrollo realizado, concluimos que un punto óptimo (dadas las restricciones actuales en lo que hace a la cantidad de observaciones) que garantiza calidad de la información aceptable y niveles de privacidad tolerables, es un presupuesto global de $4\varepsilon'$ y un δ de 0,0005 por consulta (parámetro sólo válido para el mecanismo gaussiano). Si bien

este es un punto a seguir trabajando, especialmente para ofrecer niveles de privacidad más estrictos, consideramos que ya implican una mejora respecto a los métodos actuales de distribución de la información. Asimismo, se han tenido en cuenta para dicho análisis las expectativas de privacidad de los individuos, donde se ha ponderado el hecho de que no existe una expectativa de privacidad total, dado el carácter público de los procesos y de las resoluciones.

Finalmente, hemos concluido sobre la necesidad de expandir el tamaño del conjunto de datos disponible, especialmente en lo que refiere a causas de violencia de género. Mayor cantidad de observaciones no sólo llevarán a disminuir el ruido incorporado, sino a habilitar la distribución de más información, ya sea incorporando nuevos atributos, o presentando tablas de contingencia. Otra ventaja de ampliar el número de observaciones disponibles es que permitirá ofrecer garantías de privacidad más estrictas que las actuales.

Esquema IV. Tareas y consideraciones relevantes para la implementación de la solución propuesta.



Conclusiones

En base al marco teórico expuesto, resulta evidente que la estrategia actual de publicación de los datos de parte del juzgado es insuficiente frente a las crecientes amenazas que el avance tecnológico y la virtualización de la vida social representan sobre la privacidad de los individuos involucrados. A pesar de que la publicación de los datasets estructurados cumplen con las recomendaciones básicas de privacidad, como las presentadas en las Reglas de Heredia, que sugieren excluir datos directamente identificatorios (PII) de las publicaciones on - line, tal como hemos visto, esto no protege de forma suficiente contra un ataque de identificación.

Dadas estas necesidades, nos hemos abocado al estudio de la situación actual en busca de posibles mejoras a implementar en las estrategias de publicación de los conjuntos de datos. En este sentido, consideramos que los principios de DP se ajustan a los requerimientos planteados. A pesar de ciertos obstáculos y limitaciones de índole técnica, hemos logrado implementar un flujo de trabajo capaz de atender las necesidades de privacidad y de calidad de la información. Para ello, hemos recomendado la publicación de la información en formato de tablas de distribución de frecuencias y un conjunto de transformaciones, mecanismos y parámetros para procesar la información que han sido expuestos en los respectivos apartados del trabajo.

No obstante, a pesar de estos logros, los mismos siguen representando un estadio muy primitivo del abordaje del problema, por cuanto existen puntos pendientes sobre los que debe seguirse trabajando. Futuros enfoques debieran apuntar a lograr un formato de salida más versátil, proveyendo al menos tablas de contingencia, una mejor calidad en la salida y mayores niveles de privacidad. Por otro lado, habría que contemplar la posibilidad de trabajar con conjuntos de datos que arriban en *streaming*, ya que el enfoque actual fue el de batches de datos independientes. Llegado el caso, este enfoque podría mantenerse, pero ello no quita que este problema debe analizarse en profundidad.

Adicionalmente, de cara a futuras iteraciones debe trabajarse definiendo mejor los datos recopilados por el juzgado, analizando en detalle las necesidades de los usuarios de esta información. De este modo, se podrá priorizar aquella información valiosa para el público y descartar aquella que no sea de interés. En este análisis no sólo debería considerarse la información a publicar, sino también la forma en que se organiza la misma. Deberían buscarse criterios más claros de agrupación de los datos y evitar variables con excesiva cardinalidad.

Uno de los principales emergentes a lo largo del trabajo fue la limitación de la cantidad de registros disponibles. Tal como lo señala la teoría, los mecanismos de DP son 'hambrientos' de datos. Al contar con una cantidad reducida de datos, la calidad de las salidas privadas se deteriora marcadamente. Por ello, es necesario trabajar con mayores volúmenes de datos, lo que permitirá mejorar la calidad de la salida, ensayar nuevos formatos de presentación, como tablas de contingencia, y garantizar niveles de privacidad más elevados.

En línea con lo anterior, recomendamos la búsqueda de sinergias con otros juzgados, con miras a coordinar la iniciativa de datos abiertos. De este modo, podría accederse a mayores volúmenes de datos y uniformar criterios de tratamiento y exposición de la información. Asimismo, recomendamos escalar esta iniciativa a nivel nacional, de forma que pueda integrarse a los esfuerzos encarados desde el Ministerio de Justicia en el marco del Programa de Justicia Abierta. De este modo, no sólo se lograría un corpus de registros más importante, sino que también se evitaría el solapamiento de iniciativas con el potencial de goteo de datos que ello implica.

Acordadas de los tribunales superiores de justicia referidas en el trabajo

Reglas de Heredia : Aplicación obligatoria. Tribunal Superior de Justicia de Rio Negro. ACORDADA Nº 112 de 2003. Rio Negro.

Reglas de Heredia. Tribunal Superior de Justicia de Chubut. Acuerdo Plenario 3701 de 2008. Chubut.

Leyes referidas en el trabajo

Pacto Internacional de Derechos Civiles y Políticos [PIDCP]. Ratificado en la Argentina por la ley 23.313 del año 1986. Argentina.

Ley de Protección de Datos Personales [LPDP]. Ley 25.326 de 2000. Argentina.

General Data Protection Regulation [GDPR]. Regulación 679 de 2016. Unión Europea

Health Insurance Portability and Accountability Act [HIPAA] Ley 104-191 de 1996. Estados Unidos.

Derecho de acceso a la información pública [DAIP] Ley 27.275 de 2016. Argentina.

Ley de acceso a la información pública [LAIP]. Ley 104 de 1998. Ciudad Autónoma de Buenos Aires.

Bibliografía referida en el trabajo

Abadi, Martín; Chu, Andy; Goodfellow, Ian; McMahan, H. Brendan; Mironov, Ilya; Talwar, Kunal & Zhang, Li. (2016). “*Deep learning with differential privacy*”. En *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308-318.

Abowd, Jhon. M.; Garfinkel, Simson & Martindale, Christian. (2019). “*Understanding database reconstruction attacks on public data*”. En *Communications of the ACM*, *62*(3), 46-53.

Abowd, Jhon. M.; Garfinkel, Simson & Powazek, Sarah. (2018). “*Issues encountered deploying differential privacy*”. En *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 133-137.

Apple, *Differential Privacy Team* (2017). *Learning with privacy at scale.* Consultado on - line en https://machinelearning.apple.com/research/learning-with-privacy-at-scale el 15-12-2020

Baiyere, Abayomi & Hukal, Philipp. (2020). "*Digital disruption: a conceptual clarification*". En *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 5482-5491.

Bartolucci, Mónica. (2017). "Un largo y sigiloso camino. Espionaje e infiltración policial en el mundo estudiantil en la Argentina (1957-1972)" En *Diacronie. Studi di Storia Contemporanea*, 29(1), Consultado On Line en http://journals.openedition.org/diacronie/5221 el 16-1-2021

Basterra, Marcela I. (2018) *Acceso a la información pública y transparencia*, Buenos Aires, Astrea.

Bird, Sarah; Allen, Joshua & Walker, Kathleen (2020). *A Platform for Differential Privacy*. Consultado on-line en https://www.microsoft.com/en-us/research/uploads/prod/2020/05/DPwhitepaper.pdf el 12-1-2021

Bluemke, Emma; Cuervas-Mons, Claudia G.; Trask, Andrew; Dafoe, Allan & Garfinkel, Ben. (2020). "*Beyond Privacy Trade-offs with Structured Transparency*". En *arXiv preprint arXiv:2012.08347*.

Bolot, Jean & Zang, Hui (2011). "*Anonymization of location data does not work: A large-scale measurement study*". En *Proceedings of the 17th annual international conference on Mobile computing and networking.* 145-156.

Comisión Europea. (2012). *La Comisión propone una reforma general de las normas de protección de datos para aumentar el control de los usuarios sobre sus propios datos y reducir los costes para las empresas*. Referencia IP/12/46.

Cornwall, Mark. (1992). "News, Rumour and the Control of Information in Austria-Hungary, 1914–1918" En *History*, *77*(249), 50-64.

Dalenius, Tore (1977). "*Towards a methodology for statistical disclosure control"*. En *Statistik Tidskrift,* 15, 429-444.

Desfontaines, Damien (2021). *The magic of Gaussian noise.* Consultado on-line en https://desfontain.es/privacy/gaussian-noise.html el 10-3-2021

Ding, Bolin; Kulkarni, Janardhan & Yekhanin, Sergey (2017) "*Collecting telemetry data privately"*. En *Advances in Neural Information Processing Systems* 30, NIPS '17, 3571–3580.

Dwork, Cynthia; McSherry, Frank; Nissim, Kobbi & Smith, Adam (2006) "*Calibrating Noise to Sensitivity in Private Data Analysis*" En Halevi S., Rabin T. (eds) *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science*, vol 3876. Springer, Berlin, Heidelberg.

Dwork, Cynthia, & Roth, Aaron. (2014). "*The algorithmic foundations of differential privacy*". En *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407.

Erlingsson, Ulfar; Pihur, Vasyl & Korolova, Aleksandra (2014). "*RAPPOR: Randomized aggregatable privacy-preserving ordinal response"*. En *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, 1054–1067.

Fathima, Shaistha (2020). *Query "Sensitivity" types and effects on Differential Privacy Mechanism*. Consultado on - line en https://becominghuman.ai/query-sensitivity-types-and-effects-on-differential-privacy-mechanism-c94fd14b9837 el 25-2-2021

Free Access Law Movement [FALM]. (2002). *Declaration on Free Access to Law.* Consultado on - line en http://www.falm.info/declaration/ el 1-4-2021

Gómez Zavaglia, Tristán (2020). "*Lineamientos legislativos y jurisprudenciales del acceso a la información pública*". En *Red Sociales, Revista del Departamento de Ciencias Sociales*, 07(07), 65-82.

Goodfellow, Ian & Papernot, Nicolas (2018). *Privacy and machine learning: two unexpected allies?* Consultado on-line en http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html el 5-4-2021

Jefatura de Gabinete de Ministros de Argentina (2020). *Cuarto Plan de Acción de Gobierno Abierto*. Consultado on line en https://www.opengovpartnership.org/wp-content/uploads/2019/10/Argentina_Action-Plan_2019-2022_Revised.pdf el 15-4-2021

Kamath, Gautam (2020a). *Intro to Differential Privacy, Part 2* [Material de clase del curso *Algorithms for Private Data Analysis*]. Consultado on-line en http://www.gautamkamath.com/CS860notes/lec4.pdf el 15-12-2020

Kamath, Gautam (2020b). *Approximate Differential Privacy* [Material de clase del curso *Algorithms for Private Data Analysis*]. Consultado on-line en http://www.gautamkamath.com/CS860notes/lec5.pdf el 15-12-2020

Kamath, Gautam (2020c). *Deployments of DP: Local Differential Privacy* [Material de clase del curso *Algorithms for Private Data Analysis*]. Consultado on-line en http://www.gautamkamath.com/CS860notes/lec17.pdf el 15-12-2020

Kolata, Gina. (2019). "*You Got a Brain Scan at the Hospital. Someday a Computer May Use It to Identify You*". Nota publicada en *N.Y Times* el 23/10/2019, consultada On Line en https://www.nytimes.com/2019/10/23/health/brain-scans-personal-identity.html el 20/11/2020

Kosinski, Michal; Stillwell, David & Graepel, Thore. (2013). "*Private traits and attributes are predictable from digital records of human behavior*". En *Proceedings of the national academy of sciences*, *110*(15), 5802-5805.

Li, Gang; Yu, Philip S.; Zhou, Wanlei & Zhu, Tianqing. (2017). *Differential privacy and applications*. Cham, *Springer International Publishing*.

Lippert, Christoph; Sabatini, Riccardo M.; Maher, Cyrus; Kang, Eun Yong; Lee, Seunghak; Arikan, Okan; Harley, Alena; Bernal, Axel; Garst, Peter; Lavrenko, Victor; Yocum, Ken; Wong, Theodore; Zhu, Mingfu; Yang, Wen-Yun; Chang, Chris; Lu, Tim; W. H. Lee, Charlie; Hicks, Barry; Ramakrishnan, Smriti; Tang, Haibao; Xie, Chao; Piper, Jason; Brewerton, Suzanne; Turpaz, Yaron; Telenti, Amalio; Roby, Rhonda K.; Och, Franz J. & Venter J., Craig (2017). "*Identification of individuals by trait prediction using whole-genome sequencing data*". En *Proceedings of the National Academy of Sciences*, *114*(38), 10166-10171.

McKay Bowen, Claire & Snoke, Joshua (2019). *Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge*. *arXiv:1911.12704*

McMahan, H. Brendan; Andrew, Galen; Erlingsson, Ulfar; Chien, Steve; Mironov, Ilya; Papernot, Nicolas & Kairouz , Peter (2019). *A General Approach to Adding Differential Privacy to Iterative Training Procedures*. arXiv:1812.06210

Narayanan, Arvind & Shmatikov, Vitaly. (2006). "*How to break anonymity of the netflix prize dataset*". En *Cryptography and Security*, *eprint arXiv:cs/0610105*

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2021). *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*. National Academies Press.

National Institute of Standards and Technology [NIST] (2007). *OMB Memorandum 07-16*. Consultado on-line en https://georgewbush-whitehouse.archives.gov/omb/memoranda/fy2007/m07-16.pdf

Nissenbaum, Helen. (2010). *Privacy in context*. Stanford, *Stanford Law Books*.

Nissenbaum, Helen. (2011). "*A contextual approach to privacy online"*. En *Daedalus*, 140(4), 32-48.

OEA (2009). *Ley modelo Interamericana sobre acceso a la información*. Consultada on - line en http://www.oas.org/es/sla/ddi/docs/acceso_informacion_Texto_de_Ley_editado_DDI.pdf el 10-4-2021

Peng, Han; Gong, Weikang; Beckmann, Christian F.; Vedaldi, Andrea & Smith, Stephen. M. (2021). "*Accurate brain age prediction with lightweight deep neural networks*". En *Medical Image Analysis,* 68.

Reglas de Heredia (2003). *Recomendaciones aprobadas durante el Seminario Internet y Sistema Judicial realizado en la ciudad de Heredia (Costa Rica)*, los días 8 y 9 de julio de 2003. Consultada on - line en https://archivos.juridicas.unam.mx/www/bjv/libros/4/1646/24.pdf el 20-3-2021

Wang, Yilun & Kosinski, Michal. (2018). "*Deep neural networks are more accurate than humans at detecting sexual orientation from facial images*". En *Journal of personality and social psychology*, *114*(2), 246.

Anexos

Anexo I. Descripción de los campos del dataset original (provista por el Juzgado)

N: número único dentro del Set de Datos. Nos permite conocer rápidamente cuántas resoluciones se dictaron;
NRO_REGISTRO: número de registro interno del Juzgado que lleva cada una de las resoluciones (orales y escritas);
FECHA_RESOLUCION: día de la resolución. En caso de las audiencias orales es el día de su inicio;
FIRMA: indica el/la Juez/a que firmó la resolución. En nuestro Set las opciones son Pablo C. Casas -titular del Juzgado- o Juez/a interinamente a Cargo -cuando el titular se encuentra de licencia-;
MATERIA: es la competencia del Juzgado para intervenir en los casos, Puede ser penal, contravencional, faltas, amparo, habeas corpus o ejecuciones de multa;
ART_INFRINGIDO: artículo/s de la/s infracción/es en el caso;
CODIGO_O_LEY: referido a la categoría anterior, indica si el artículo pertenece al Código Penal de la Nación, Código Contravencional, Ley 451 - Régimen de Faltas de la Ciudad de Buenos Aires, Ley 23737 - Tenencia y tráfico de estupefacientes, Ley 13944 - Incumplimiento de los deberes de asistencia familiar, Ley 24720 - Impedimento de contacto con padre no conviviente, Ley 14346 - Malos tratos o actos de crueldad a los animales, Ley 26735 - Régimen Penal Tributario, Ley 12331 - Ley Nacional de Profilaxis;
CONDUCTA: se indica la acción relativa al delito, la contravención, o la falta que aparece descripta en el artículo infringido;
CONDUCTA_DESCRIPCION: relacionado con el punto 10) donde especificamos si el delito, la contravención o la falta infringida tiene alguna particularidad, como puede ser que se encuentra agravada por alguna causal;
VIOLENCIA_DE_GENERO: si el hecho objeto de investigación se encuentra dentro de un contexto de violencia de género o no;
V_FISICA: indica si por la declaración de la víctima hubo violencia física.
En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.
La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia física, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;
V_PSIC: indica si por la declaración de la víctima hubo violencia psicológica;
En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia psicológica, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

V_ECON: indica si por la declaración de la víctima hubo violencia económica o patrimonial.

En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia económica, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

V_SEX: indica si por la declaración de la víctima hubo violencia sexual.

En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia sexual, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

V_SOC: indica si por la declaración de la víctima hubo violencia social.

En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia social, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

V_AMB: indica si por la declaración de la víctima hubo violencia ambiental.

En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia ambiental, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

V_SIMB: indica si por la declaración de la víctima hubo violencia simbólica.

En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia simbólica, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

V_POLIT: indica si por la declaración de la víctima hubo violencia política.

En el set de datos las opciones son “si”, “no”, “s/d” (sin datos), y “no_corresponde”.

La diferencia entre las variables “no” y “no_corresponde”, es que la primera de ellas indica que en un caso con contexto de violencia de género no hubo violencia simbólica, y la otra se utiliza cuando en el caso no hay un contexto de violencia de género;

FRASES_AGRESION: transcripción de las frases descriptas por la víctima como la agresión verbal sufrida y que son parte de los hechos del caso. Aplica para los casos de violencia verbal;

MODALIDAD_DE_LA_VIOLENCIA: es la forma en que se manifiestan los distintos tipos de violencia. Las opciones dentro del set son: doméstica, institucional, mediática, laboral, contra la libertad reproductiva, obstétrica, en espacio público o privado, y política y pública;

GÉNERO ACUSADO/A: indica el género de la persona acusada;

NACIONALIDAD_ACUSADO/A: indica la nacionalidad de la persona acusada;

EDAD_ACUSADO/A_AL_MOMENTO_DEL_HECHO: indica la edad de la persona acusada al momento del hecho;

NIVEL_DE_INSTRUCCION_ACUSADO/A: nivel de estudios formales alcanzados por la persona acusada;

GENERO_DENUNCIANTE: indica el género de la persona que denuncia;

NACIONALIDAD_DENUNCIANTE: indica la nacionalidad de la persona que denuncia;

EDAD_DENUNCIANTE_AL_MOMENTO_DEL_HECHO: indica la edad de la persona que denuncia al momento del hecho;

NIVEL_DE_INSTRUCCION_DENUNCIANTE: indica los estudios cursados por la persona que denuncia;

FRECUENCIA_EPISODIOS: se indica carácter esporádico (cuando las agresiones ocurren de forma aislada), diario, habitual (si ocurren semanalmente), eventual (si se dan quincenal o mensualmente), o si se trata de la primera agresión sufrida;

RELACION_Y_TIPO_ENTRE_ACUSADO/A_Y_DENUNCIANTE: indica el tipo de vínculo que tiene la persona acusada con la denunciante;

HIJOS_HIJAS_EN_COMUN: indica si la persona acusada y la denunciante tienen hijos/as en común;

MEDIDAS_DE_PROTECCION_VIGENTES_AL_MOMENTO_DEL_HECHO: indica las medidas de protección que se hayan impuesto para proteger a la víctima y si estaban vigentes al momento de los hechos;

ZONA_DEL_HECHO: indica la zona en la que sucedió el hecho;

LUGAR_DEL_HECHO: indica lugar físico (ambiente o vía pública) donde ocurrieron los hechos o si fue cometido mediante medios tecnológicos;

TIPO_DE_RESOLUCION: interlocutorias son aquellas que definen una cuestión concreta durante la tramitación del proceso, o definitivas son aquellas resoluciones que ponen fin al proceso de la causa;

OBJETO_DE_LA_RESOLUCION: sobre qué se resolvió;

DETALLE: especifica el punto anterior respecto a qué se resolvió;

DECISION: si hace lugar o no al planteo realizado por las partes;

ORAL_ESCRITA: indica si la resolución fue dictada en audiencia -oral- o no -escrita;

HORA_DE_INICIO : horario de inicio de la audiencia;

HORA_DE_CIERRE: horario de finalización de la audiencia;

LINK: link de acceso a la resolución en formato abierto;

DURACION: indica el tiempo que insumió la realización de la audiencia;

SI_NO_RECURRENTE: si la resolución fue apelada (cuestionada) y por cual parte del proceso (Fiscalía, Defensoría, Defensor particular, o ambos, perito, intérprete, querella, infractor/a, o contienda (este último caso refiere a cuando hay un conflicto de competencia entre dos juzgados del fuero));

DECISION_CAMARA_DE_APELACIONES: indica que resolvió la Cámara de Apelaciones del Fuero;

N_DE_REGISTRO_Y_TOMO_CAMARA: número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones y el tomo donde las encontramos;

LINK_CAMARA: link de acceso a la resolución de la Cámara de Apelaciones del Fuero;

SI_NO_RECURRENTE_CAMARA: si la resolución fue recurrida y por cual parte del proceso (Fiscalía, Defensoría, Defensor particular, o ambos);

DECISION_DE_ADMISIBILIDAD_CAMARA: si admite o no el recurso interpuesto;

N_DE_REGISTRO_Y_TOMO_CAMARA_1: número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones y el tomo donde las encontramos;

LINK_CAMARA_1: acceso a la resolución de la Cámara de Apelaciones del Fuero;

QUEJA_Y_RECURRENTE: quien pide que se eleve al Tribunal Superior de Justicia (en adelante TSJ);

DECISION_DE_ADMISIBILIDAD_TSJ: si se admite la intervención del TSJ;

N_DE_REGISTRO_Y_TOMO_TSJ: número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones del fuero;

LINK_TSJ: número de registro interno del Juzgado para las resoluciones de la Cámara de Apelaciones y el tomo donde las encontramos;

DECISION_DE_FONDO_TSJ: si se admitió, qué decidió el TSJ;

N_DE_REGISTRO_Y_TOMO_TSJ_1: número de registro interno del Juzgado para las resoluciones del Tribunal Superior de Justicia;

LINK_TSJ_1: acceso a la resolución del Tribunal Superior de Justicia.

RECURSO_EXTRAORDINARIO_Y_RECURRENTE: indica si se presentó un recurso para llegar a la Corte Suprema de Justicia de la Nación (en adelante CSJN);

DECISION_CSJN: indica que decidió la CSJN;

N_REGISTRO_Y_TOMO_CSJN: de la resolución de la CSJN;

LINK_CSJN: acceso a la resolución de la CSJN.

Anexo II. Procedimiento para la aplicación de los mecanismos de DP y el cálculo de las métricas de error.

1. Se aplica el mecanismo de DP sobre cada atributo del dataset.
2. Para cada atributo del dataset, se calcula el error entre la distribución de probabilidades original y la privada.
3. Se repiten los pasos 1 y 2 n veces, para garantizar resultados más robustos.
4. Se promedian las métricas de error para cada atributo.
5. Se promedian las métricas de error calculadas en el paso 4.

Anexo III. Pseudocódigo de los mecanismos laplace truncado y sin truncar

Implementación del mecanismo laplaciano truncado

```
nueva_dist_prob = [ ]  array para almacenar la nueva distribución de probabilidades
for i, p in enumerate(distr_prob_orig):
      ruido =  Lap(escala)  toma aleatoriamente un valor de esta distribución
      nueva_dist_prob [i] = p + ruido
      if nueva_dist_prob [i]< 0:
            nueva_dist_prob [i]= 0
nueva_dist_prob = nueva_dist_prob / sum(nueva_dist_prob)  normalizo la salida
```

Implementación del mecanismo laplaciano sin truncar

```
nueva_dist_prob = [ ]  array para almacenar la nueva distribución de probabilidades
for i, p in enumerate(distr_prob_orig):
      ruido =  Lap(escala)  toma aleatoriamente un valor de esta distribución
      nueva_dist_prob [i] = p + ruido
sumo densidad negativa para redistribuir
dens_bajo_cero = - sum(nueva_dist_prob[nueva_dist_prob<0])
convierto las probabilidades negativas en 0
nueva_dist_prob[nueva_dist_prob<0]=0
asigno proporcionalmente la densidad negativa sobre el resto del dominio de la distribución
redistribuir = dens_bajo_cero * (nueva_dist_prob / sum(nueva_dist_prob))
sumo la densidad negativa al resto de la distribución
nueva_dist_prob += redistribuir
nueva_dist_prob = nueva_dist_prob / sum(nueva_dist_prob)  normalizo la salida
```
